BioSense Platform Data Flow

Part 1 – Data Ingestion into the BioSense Platform

**Division of Health Informatics and Surveillance
(DHIS) Team**

February 11, 2016

Agenda

1. **Introductions**
2. **BioSense Platform Update**
3. **Defining Data Flow**
4. **Ingesting Syndromic Surveillance Data to the BioSense Platform**
5. **BioSense Platform Archive: Data Element Requirements**
6. **Next Steps**

Speakers

- **Mike Coletta, CDC**
- **Roseanne English, CDC**

- **Max Worlund, Contractor**
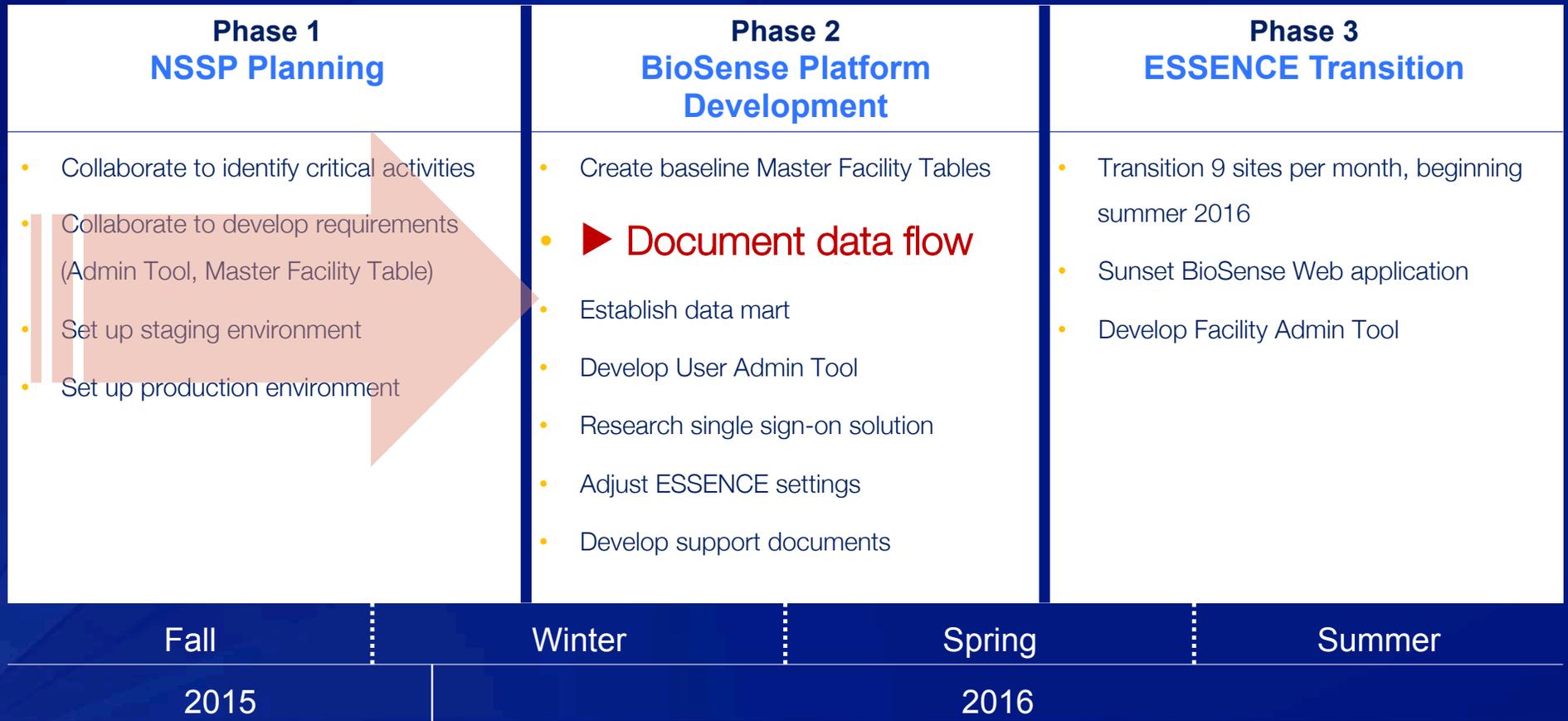- **Charlie Ishikawa, Contractor**
- **Lindsay Brown, Contractor**

BIOSENSE PLATFORM UPDATE FEBRUARY 2016

Mike Coletta, CDC

Public Health Syndromic Surveillance Data Flow

BioSense Platform Update

Phase 1 **NSSP Planning**	**Phase 2** **BioSense Platform Development**	**Phase 3** **ESSENCE Transition**
• Collaborate to identify critical activities • Collaborate to develop requirements (Admin Tool, Master Facility Table) • Set up staging environment • Set up production environment	• Create baseline Master Facility Tables • ▶ Document data flow • Establish data mart • Develop User Admin Tool • Research single sign-on solution • Adjust ESSENCE settings • Develop support documents	• Transition 9 sites per month, beginning summer 2016 • Sunset BioSense Web application • Develop Facility Admin Tool

Fall	Winter	Spring	Summer
2015		2016	

BIOSENSE PLATFORM
DATA FLOW REQUIREMENTS

Roseanne English, CDC

BioSense Platform Data Flow Goals

- **Enable public health officials to conduct syndromic surveillance more systematically and collaboratively**

- **Provide access to data delivered to the BioSense Platform**
  - HL7 messages
  - ASCII converted to HL7 messages

- **Provide access to stable structured databases reflecting processed HL7 messages**

- **Prime the data with various calculated fields to allow for flexibility in analysis approaches regardless of the analytic tool being used**

- **Make the data complete, useful, and actionable**

BioSense Platform Data Flow Design

- **Ingest syndromic surveillance data**
- **Remove Personally Identifiable Information (PII) from targeted message segments**
  - HL7 requires that certain fields are populated even when data should not be sent
  - The BioSense Platform will scrub incoming data to remove PII from these segments
- **Separate servers for analytic data marts to optimize transactional data processing**
- **Store syndromic surveillance data to support analyses**
  - Include all core data elements for syndromic surveillance
  - Include additional standardized data elements to support analytics
  - Map and package incoming data for ESSENCE interactive application

Technology Considerations
- Mirth and PERL processing for data ingestion and transformation
- Microsoft SQL Server for data storage
- ESSENCE application
- Other query and analysis tools (e.g., R, SAS®)
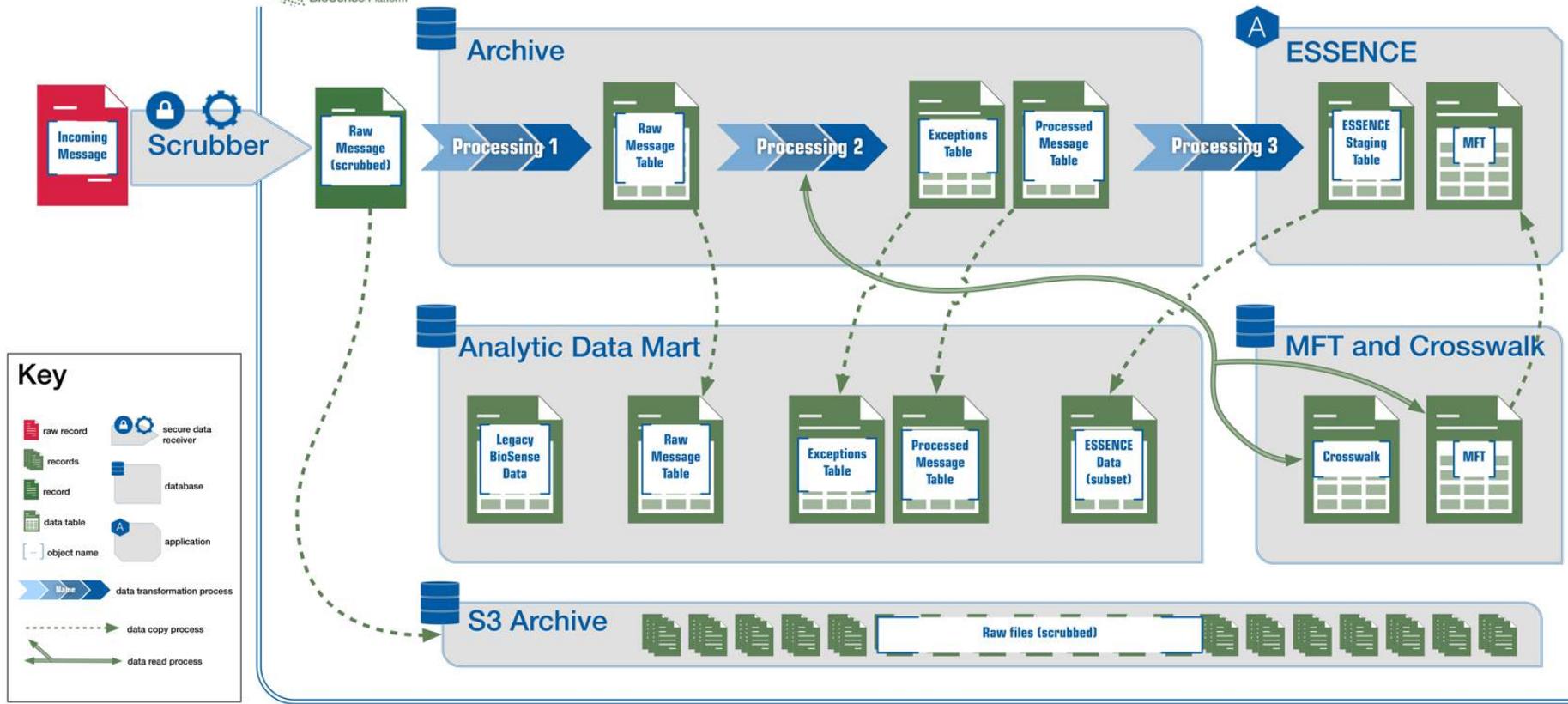
BioSense Platform Data Flow

Process Components:

1. "Scrub" targeted HL7 message components / incoming data to remove PII

2. Store archival copies of incoming data files

3. Ingest data into a BioSense Platform Archive

4. Ingest data into the ESSENCE application

5. Populate an analytic data mart to support complex analyses

6. Maintain Master Facility Table data and Crosswalk information

Last modified: January 27, 2016

NOTE: CONCEPTUAL DIAGRAM. FOR PLANNING & DISCUSSION USE ONLY.
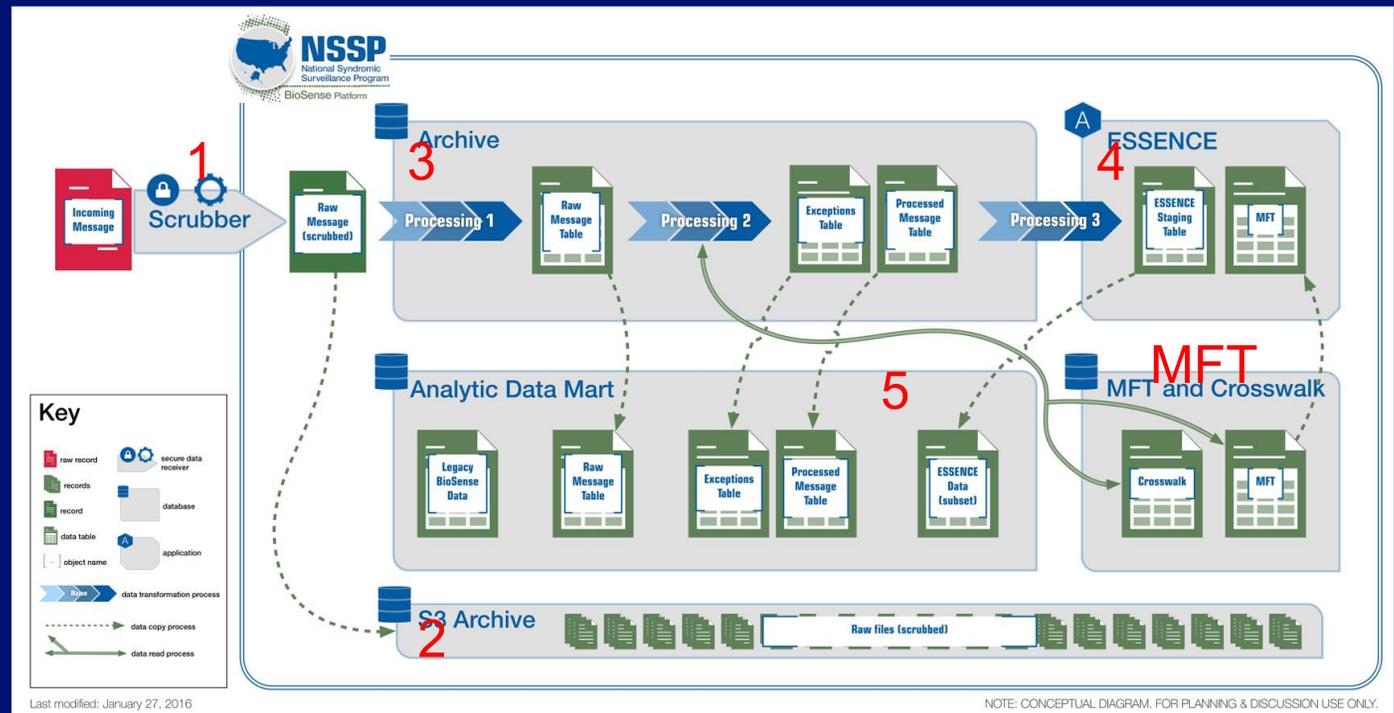
Capture details to assess timeliness of reporting

Collect date/time of key processing points

1. Time of landing on server via FTP/PHIN/etc

2. Time of ingestion to Raw table

3. Time of ingestion to Processed table

4. Time of ingestion into ESSENCE

5. Time of ingestion into the data mart
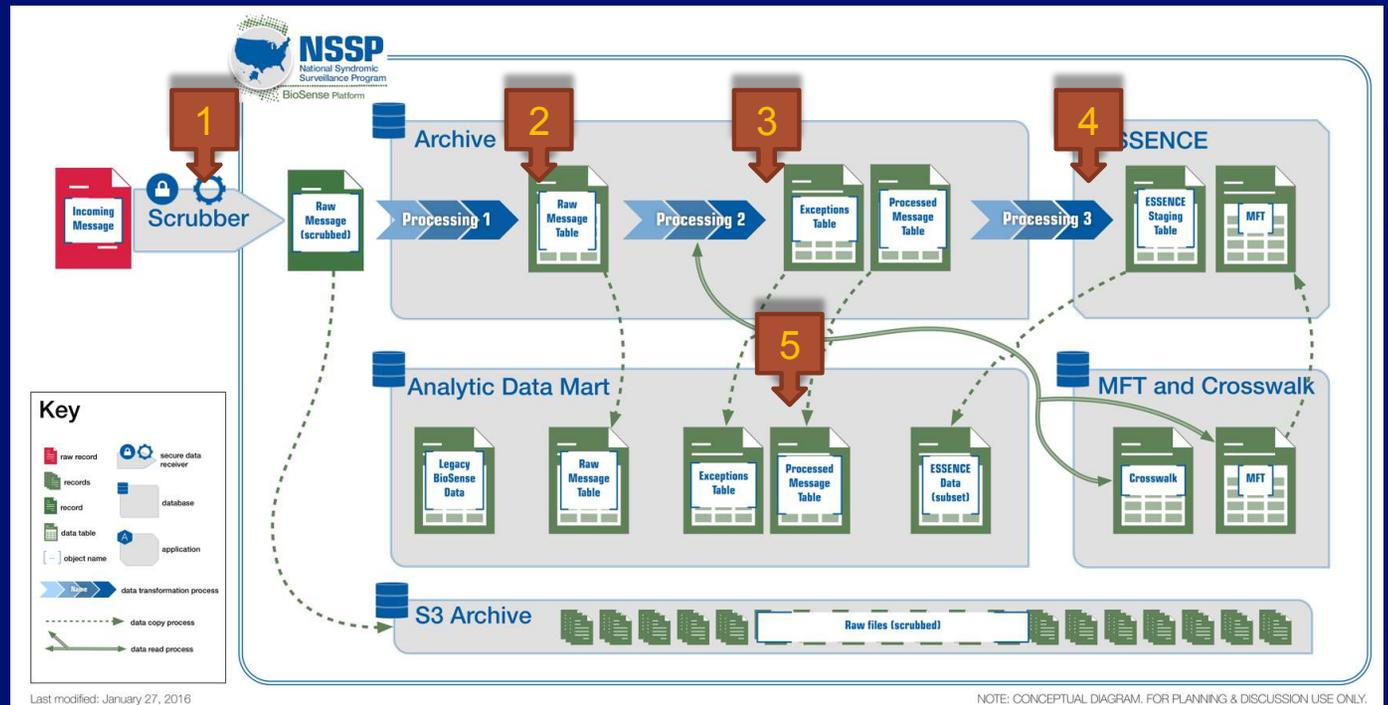
INGESTING SYNDROMIC SURVEILLANCE DATA INTO THE BIOSENSE PLATFORM

Back-end automation for target outcomes

# "Scrub": Remove Personally Identifiable Information (PII)

Challenge:

Although the PHIN guide requests that PII not be sent, incoming data may inadvertently include PII due to HL7 message requirements (e.g., "Patient Name")

Solution:

Remove PII from targeted message segments (full list on next slide)

- All incoming data will be scrubbed before processing and storage
- All data servers are encrypted
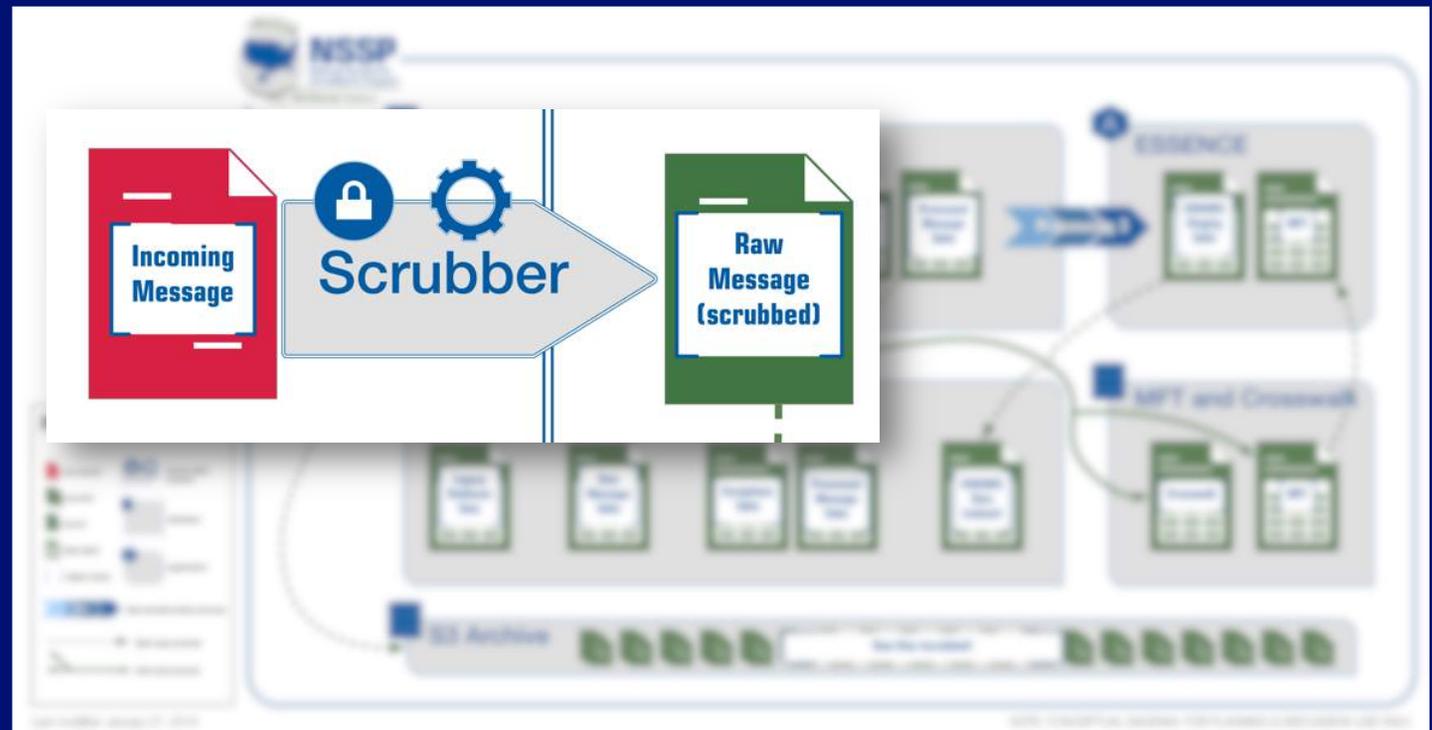
Scrubbed Segments

- PID.5.1 - Family Name
- PID.5.2 - Given Name
- PID.5.3 - Middle Name or Initial
- PID.5.4 - Suffix
- PID.5.5 - Prefix
- PID.5.6 - Degree
- PID.5.8 - Name Representation Code
- PID.5.9 - Name Context
- PID.5.10 - Name Validity Range
- PID.5.11 - Name Assembly Order
- PID.5.12 - Effective Date
- PID.6 - Mother's Maiden Name (All components)
- PID.9 - Patient Alias
- PID.11.1 - Patient Street Address Line 1
- PID.11.2 - Patient Street Address Line 2
- PID.11.8 - Patient Address - Other Geographical Designation

- PID.13 - Patient Phone Number - Home
- PID.14 - Patient Phone Number - Business
- PID.15 - Primary Language
- PID.16 - Marital Status
- PID.17 - Religion
- PID.19 - SSN Number - Patient
- PID.20 - Drivers License Number - Patient
- PID.21 - Mother Identifier
- PID.23 – Birth Place
- PID.24 - Multiple Birth Indicator
- PID.25 - Birth Order
- PID.26 - Citizenship
- PID.27 - Veterans Military Status
- PID.28 - Nationality
- PID.30.2 - Patient Death Indicator (unknown component)
- NK1.2 - Next of Kin
- NK1.4 - Next of Kin Address

- NK1.5 - Next of Kin Phone Number
- NK1.6 - Next of Kin Business Phone Number
- NK1.12 - Next of Kin Employee Number

- MRG.7 - Merge Patient - Prior Patient Name

- IN1.16 - Insured's Name
- IN1.19 - Insured's Address
- IN1.2 - Insurance Plan ID

- GT1.3-6 - Guarantor Name, Address, Phone Number
- GT1.12 - Guarantor SSN
- GT1.19 - Guarantor Employee ID Number

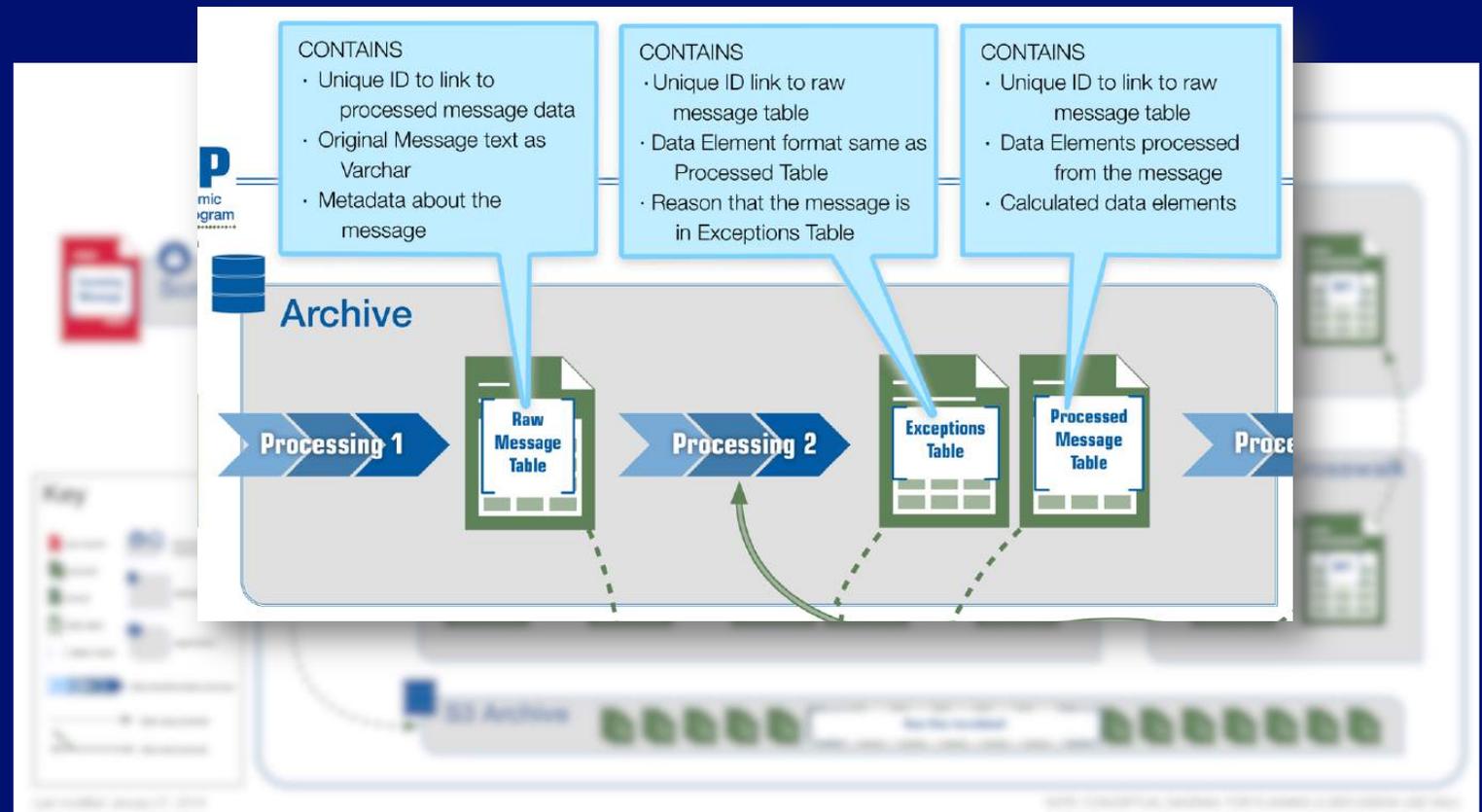# Archive - Copy Scrubbed File into S3 Archive

Scrubbed files are retained in the S3 Archive as a back up
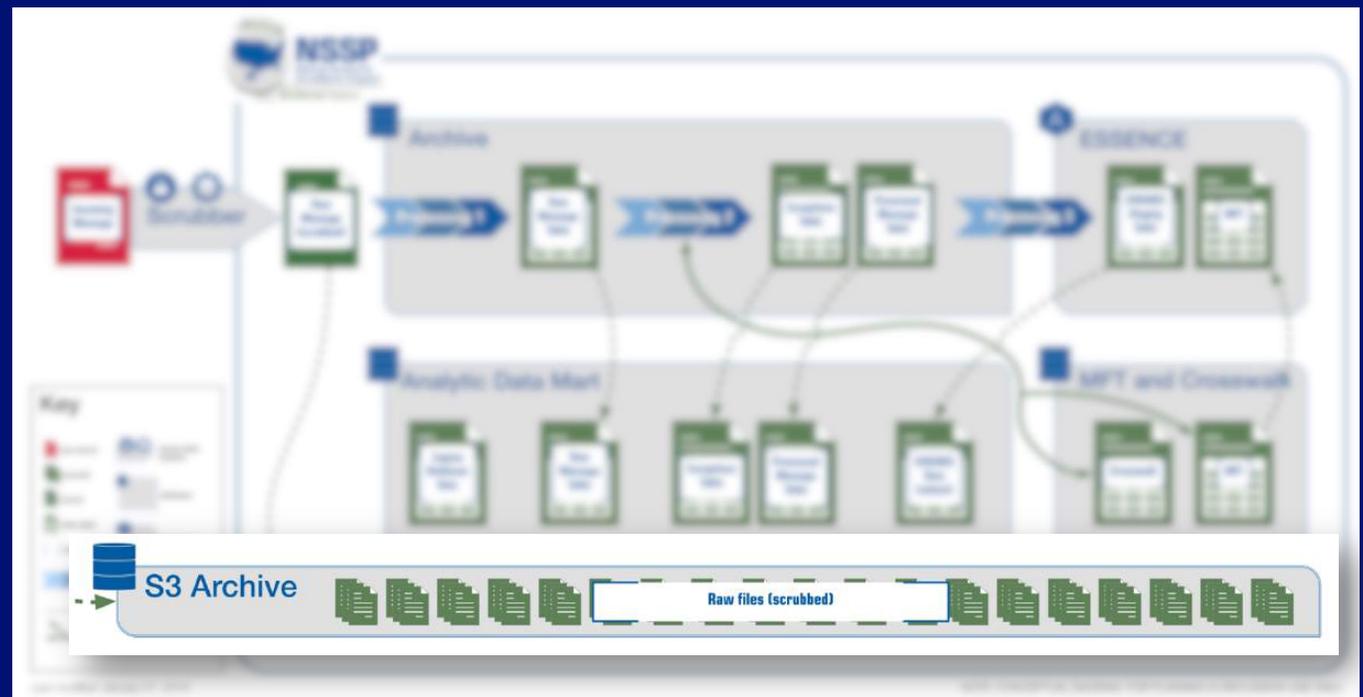
Processing – Transform data for storage in the BioSense Platform Archive

Design Considerations

- One database table per site to allow for faster processing
- PHINMS database converts and "splits" to site specific table
- Data will be indexed to optimize processing

CONTAINS
- Unique ID to link to processed message data
- Original Message text as Varchar
- Metadata about the message

CONTAINS
- Unique ID link to raw message table
- Data Element format same as Processed Table
- Reason that the message is in Exceptions Table

CONTAINS
- Unique ID to link to raw message table
- Data Elements processed from the message
- Calculated data elements

Archive

Processing 1 → Raw Message Table → Processing 2 → Exceptions Table → Processed Message Table → Proce...

BioSense Platform Archive populates ESSENCE Application

A subset of fields from the BioSense Platform Archive will be formatted, transformed (if necessary) and written to the ESSENCE application
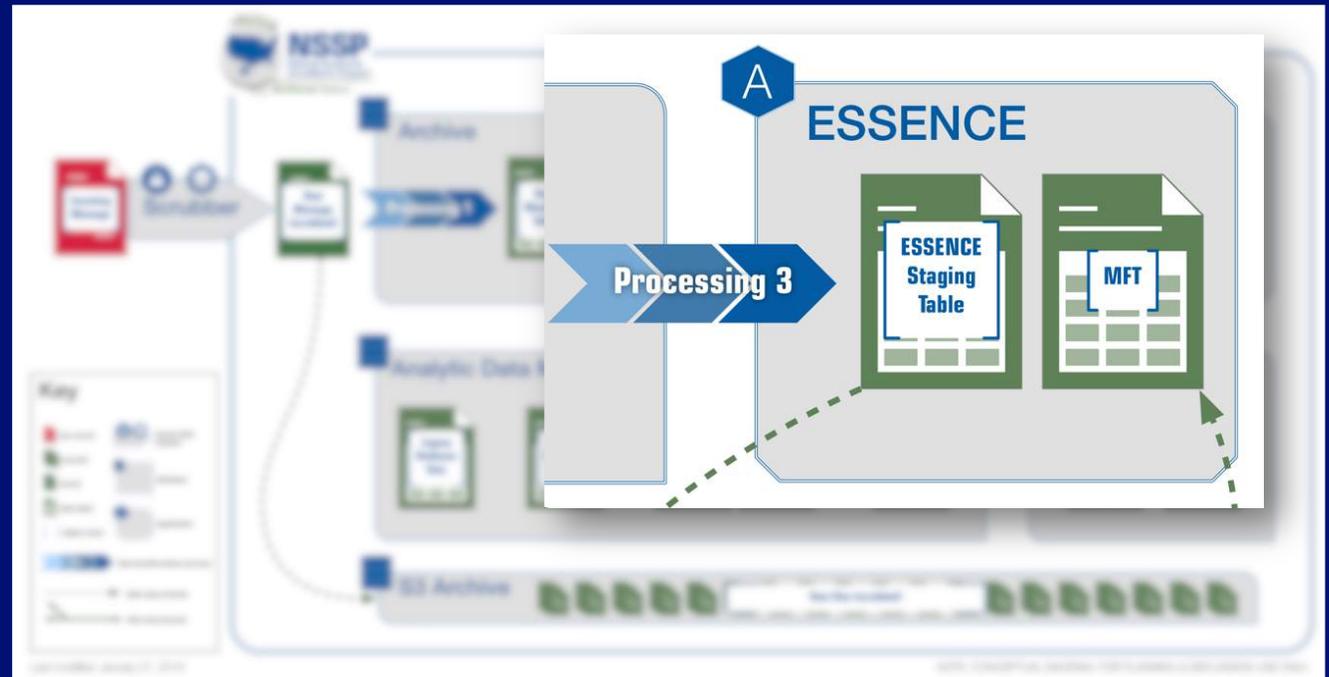
To be covered in "Part 2" in the Data Flow Webinar Series (Coming Soon)

Populating Analytic Data Mart – Supporting Complex Analytics

Initially….

- Incrementally updated copy of
  - Archive Raw Message Table
  - Archive Processed Message Table
  - Archive Exceptions Table
- Legacy BioSense data including Stage 1 and MUB
- Select ESSENCE datasets
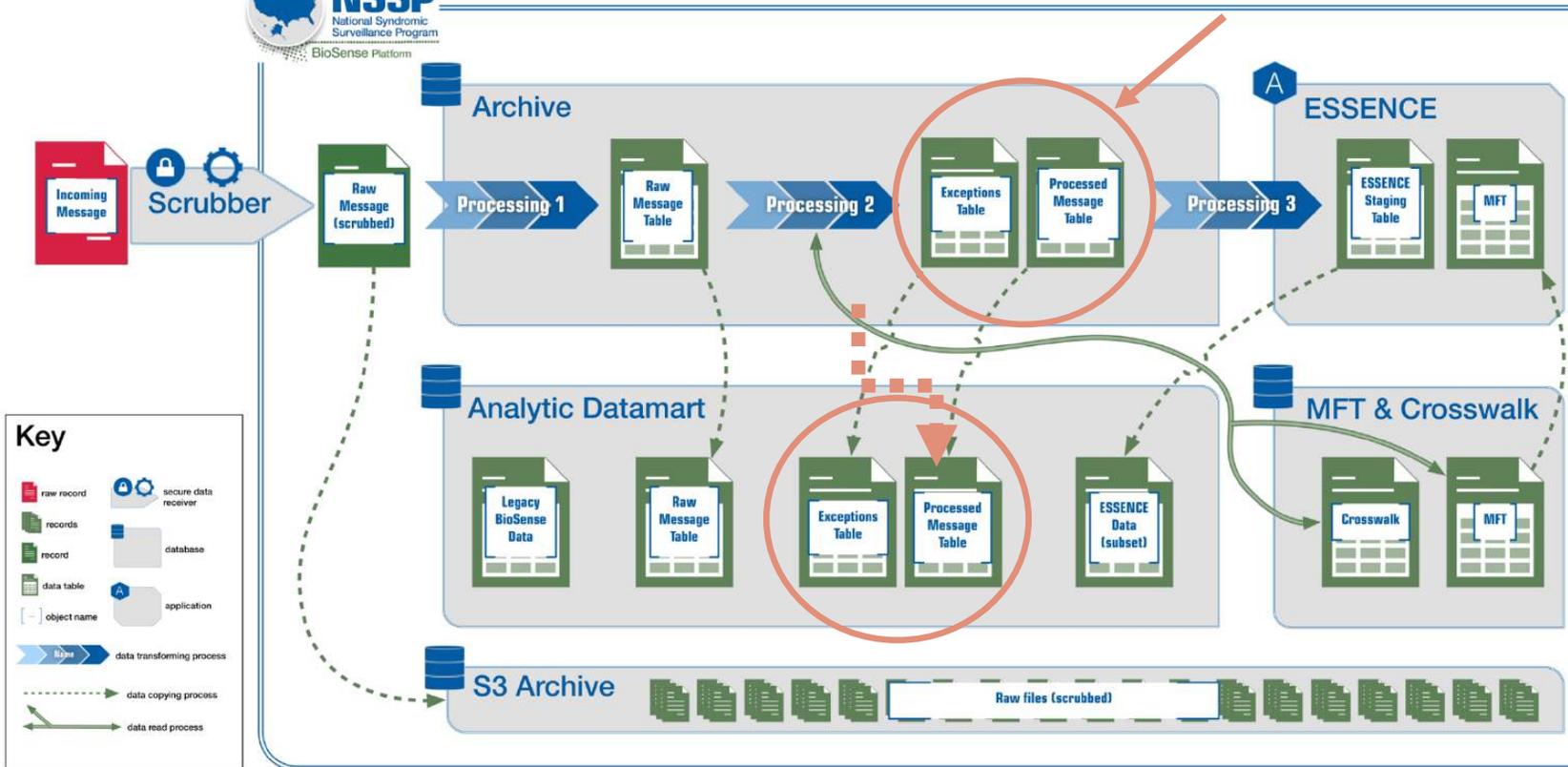
Potential additions….

- Legacy BioSense data migrated to HL7 Archive Processed Message Table – TBD based on user input
- Other marts - TBD based on user input
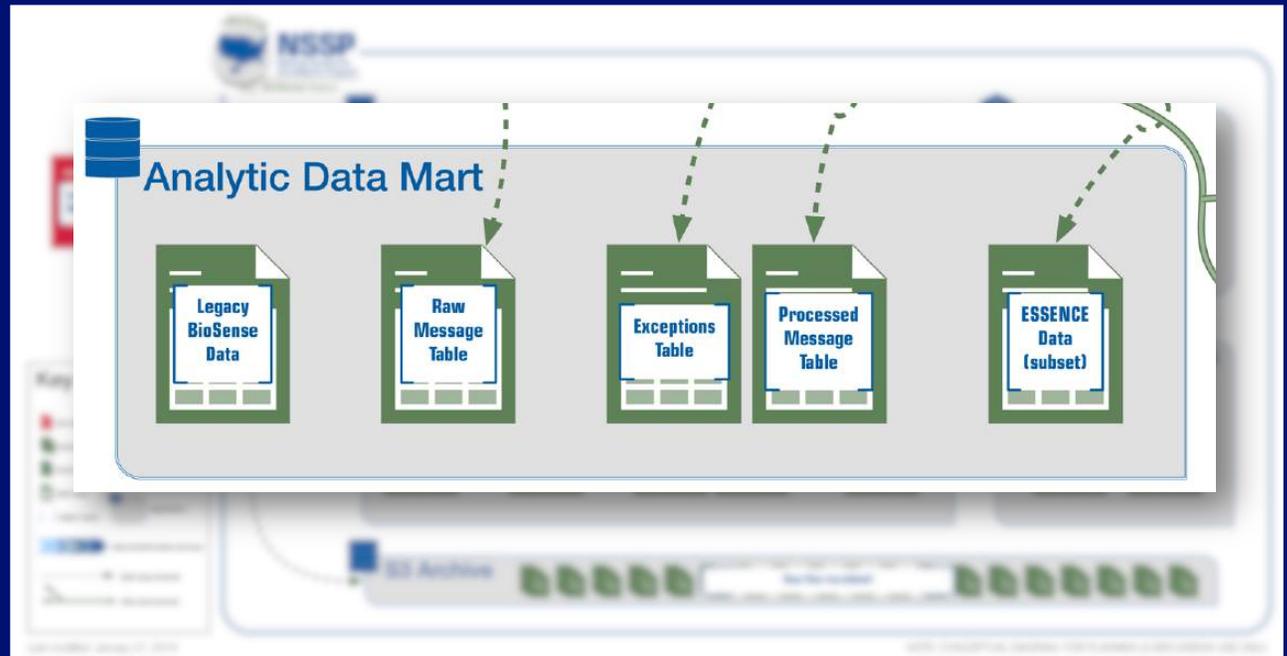
DEEP-DIVE: BIOSENSE PLATFORM ARCHIVE DATA ELEMENTS

Standardized data elements to support integration with BioSense Platform Applications

High Level Data Flow Design

Guiding Principles for Defining Data Elements

- **Load all syndromic surveillance core data elements from the PHIN Messaging Guide for Syndromic Surveillance: Emergency Department, Urgent Care, Inpatient and Ambulatory Care Settings, Release 2.0 (April, 2015)**
  - Legacy guides were considered when defining processing and incorporated where appropriate
- **Enhance analytic capability**
  - Identify new elements from a syndromic surveillance message
  - Define new processed elements
- **Ensure data elements follow a consistent naming convention**
- **Store all instances of repeating data (repeating segments or repeating fields)**
  - Separator proposed is a pipe separator: |

BioSense Platform Archive Processing Guidelines

- **If an element is a coded element (CE) or coded with exceptions (CWE) type:**
  - Store the codes in a column named *X_Code*
  - Store the code descriptions in a column named C_*Description*
  - Store the entire segment in a column named *X_Segment*

BioSense Platform Archive Processing Guidelines

If an element is calculated/ created during processing

- Store contributing elements separately
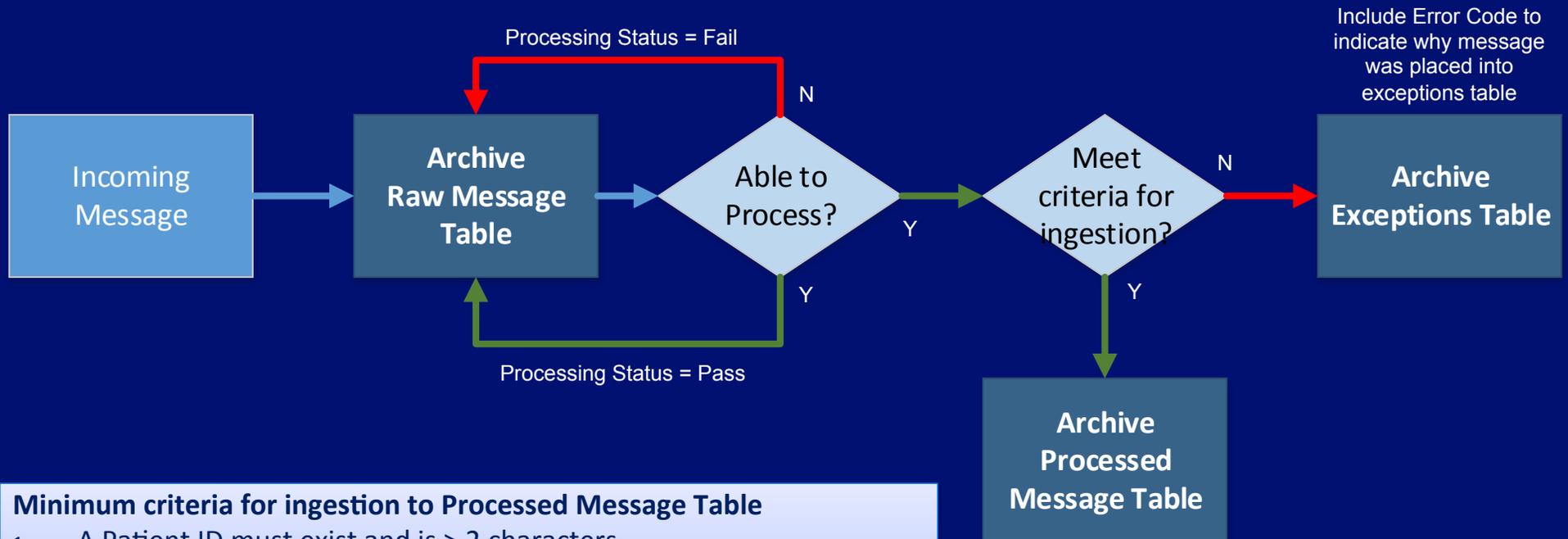- Store the field(s) contributing to the calculated value

Example:

The variable **C_Var** is calculated based on variables **Var1, Var2, Var3**

Element	Value	Logic
Var1	Null	Direct Input
Var2	A	Direct Input
Var3	B	Direct Input
C_Var	A	Select first non-null value from Var1, Var2 and Var3
C_Var_Source	Var2	Source of C_Var

Processing Exceptions

Incoming Message → Archive Raw Message Table → Able to Process? → Meet criteria for ingestion? → Archive Exceptions Table

Processing Status = Fail (N)

Processing Status = Pass (Y)

Able to Process? — Y → Meet criteria for ingestion?

Meet criteria for ingestion? — N → Archive Exceptions Table

Include Error Code to indicate why message was placed into exceptions table

Meet criteria for ingestion? — Y → Archive Processed Message Table

Minimum criteria for ingestion to Processed Message Table
1. A Patient ID must exist and is > 2 characters
2. A Visit Date must exist as a valid date and is >= 8 characters
3. A Facility ID must exist (must contain any value in EVN-7 or MSH-4)
4. The Facility ID from the message must appear as an active facility on the MFT

BioSense Platform Archive Data Elements

- **Data Element Requirements Spreadsheet**
  - Contains the full list of proposed data elements and processing for the HL7 Archive
- **Complex Data Element Requirements**
  - Chief Complaint
  - Calculated Chief Complaint
  - Calculated Patient Age
  - Calculated Death Indicator
  - Calculated Patient Class
  - Calculated Facility ID (Sending or Treating)
  - Processed Facility ID
  - Unique Patient ID & Medical Record Number
  - Visit Date/Time & Visit Date
  - Unique Visit ID
  - Processed Visit ID

Chief Complaint

- **Incoming data for Chief Complaint may be**
  - Text (TX)
  - Coded With Exceptions (CWE)

- **BioSense Platform will ingest both into a standardized set of columns**

Incoming Data is	Chief_Complaint_Code	Chief_Complaint_Text	Chief_Complaint_Segment	Chief_Complaint_Type
TX	Null	Concatenate non-null OBX-5.1	Capture the entire segment with chief complaint data	Capture the type of chief complaint data (CWE, TX, or CWE/TX)
CWE	Concatenate first non-null coded values (OBX-5.1 and OBX-5.4)	Concatenate first non-null descriptive/text values (OBX-5.2, OBX-5.5, OBX-5.9)		

Example: Chief Complaint

TX Chief Complaint: OBX|3|TX|8661-1^CHIEF COMPLAINT – REPORTED^LN||STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER|
CWE Chief Complaint: OBX|3|CWE|8661-1^CHIEF COMPLAINT:FIND:PT:PATIENT:NOM:REPORTED^LN||1941^^^^^^^^STOMACH ACHE||

Received Chief Complaints	Chief_Complaint_Code	Chief_Complaint_Text	Chief_Complaint_Segment	Chief_Complaint_Type
TX	NULL	STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER	STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER	TX
CWE	1941	STOMACH ACHE	1941^^^^^^^^STOMACH ACHE	CWE
TX and CWE	1941	STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER| STOMACHE ACHE	STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER| 1941^^^^^^^^STOMACH ACHE	CWE/TX

Calculated Chief Complaint

Combines information across message segments to reflect why a patient sought care

Fields	Description	Include?
Admit Reason Code	Short description of the provider's reason for admitting the patient	Yes
Admit Reason Description		Yes
Chief Complaint Code	Patient's self-reported chief complaint or reason for visit	Yes
Chief Complaint Text		Yes
Triage Notes	Triage notes for patient visit	No
Clinical Impression	Clinical impression (free text) of the diagnosis	Yes
Medication History	Medications prescribed/dispensed (codes or narrative)	?
Problem List	List of patient's conditions - contains a list of all complaints or clinical findings.	?
Initial Evaluation Note	Initial evaluation note from physician	?

Example: Calculated Chief Complaint

Field	Value
Admit Reason Code	78907
Admit Reason Description	ABDOMINAL PAIN, GENERALIZED
Chief Complaint Code	NULL
Chief Complaint Text	STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER
Triage Notes	Triage Notes
Clinical Impression	Pain consist with appendicitis
Calculated Chief Complaint – Current Processing	78907:SEP:ABDOMINAL PAIN, GENERALIZED:SEP:STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER:SEP:TriageNotes:SEP:Pain consist with appendicits
Calculated Chief Complaint – New Processing (proposed)	78907|ABDOMINAL PAIN, GENERALIZED|STOMACH ACHE THAT HAS LASTED 2 DAYS; NAUSEA AND VOMITING; MAYBE A FEVER|Pain consist with appendicitis

Calculated Patient Age

- The BioSense Platform receives age information inconsistently across messages
- The BioSense Platform will generate a calculated age (and companion units) by selecting the first non-null option from:
  - Calculated Visit Date – Birth Date (PID-7)
    - Default units:
      - Years (if >= 2 years)
      - Days (if < 2 years)
  - Age Reported, including units (OBX-5, OBX-6)
  - Age Calculated, including units (OBX-5, OBX-6)

Calculated Visit Date:
Return the earliest date from the following segments:
* Admit Date/Time (PV1.44.1)
* Discharge Date/Time (PV1.45.1)
* Procedure Date/Time (PR1.5.1)
* Patient Death Date/Time (PID.29.1)
* Recorded Date/Time of Message (EVN.7.2)
* Date/Time of Message (MSH.7.1)

Calculated Death Indicator

- Field is derived from several HL7 message components
- If any of the following are True then this field will be set to "Yes"
  - PID-30 (Patient Death Indicator) = "yes"
  - PID-29 (Patient Death Date/Time) is not null
  - PV1-36 (Discharge Disposition) contains "20," "40," "41," or "42"

Patient_Death_ Indicator	Death_Date_Time	Discharge_Disposition	C_Death	C_Death_Source
Yes	1/1/2016	.	Yes	Indicator
.	1/6/2016	20	Yes	DeathDate
.	.	41	Yes	Disposition
No	.	19	No	.
.	.	.	No	.

Calculated Patient Class

- Patient class is important for analyses / data explorations but may not be present in all message submissions

- BioSense Platform processing creates a Calculated Patient Class using the following algorithm:

  1) Patient Class (PV1-2) if it is a valid code for syndromic surveillance (E, I, or O)
  2) Look up the Inferred Patient Class from the Facility Type reported in the message if it is valid for syndromic surveillance
  3) Assign class value based on the inferred patient class associated with the primary entry on the MFT

Concept Code	Concept Name	Value
261QE0002X	Emergency Care [Ambulatory Health Care Facilities\Clinic/Center]	E
1021-5	Inpatient practice setting	I
261QM2500X	Medical Specialty [Ambulatory Health Care Facilities\Clinic/Center]	O
261QP2300X	Primary Care [Ambulatory Health Care Facilities\Clinic/Center]	O
261QU0200X	Urgent Care [Ambulatory Health Care Facilities\Clinic/Center]	E

Example: Calculated Patient Class

Archive Table

C_Facility_ID	Patient_Class_Code	Facility_Type_Code	C_FacType_Patient_Class	C_MFT_Patient_Class	C_Patient_Class	C_Patient_Class_Source
00010001	E	261QM2500X	O	E	E	PV1
00011114	.	261QM2500X	O	E	O	OBX
00011114	.	.	.	E	E	MFT

MFT

C_Facility_ID	Facility Type	Inferred Patient Class	Primary?
00010001	Emergency Care	E	Y
00011114	Emergency Care	E	Y
00011114	Outpatient	O	N

BioSense Platform Facility IDs

- HL7 Messages include information for both sending and treating facilities
  - BioSense Platform Archive will store both

- Facility IDs in messages may be:
  - Deprecated / legacy -> must be mapped to current ID values
  - Differ between sending / treating -> map to a "winning" ID value

Element Name	Description	Processing
Sending_Facility_ID	Sending Facility ID	Return first non-null value from: MSH-4.2 MSH-4.1
Sending_Facility_ID_Source	Source of the Sending Facility ID	Stores the source
Treating_Facility_ID	Treating Facility ID	EVN-7.2 ** Should this store EVN-7.1 if EVN-7.2 is null?**

BioSense Platform Processed Facility IDs

- Facility IDs in messages may be:
  - Deprecated / legacy -> map to current ID values through the MFT
  - Differ between sending / treating -> map to a "winning" ID value through the MFT

Element Name	Description	Processing
C_Facility_ID	Calculated Facility ID validated through MFT	Select the first valid facility ID from the MFT and pre-pend site ID: • Treating Facility • Sending Facility A valid facility ID is one that appears on the MFT. If no valid facility ID found, set to first non-null.
C_Facility_ID_Source	Source of the Calculated Facility ID	Stores the source (EVN-7.2, MSH-4.2, MSH-4.1)
C_Processed_Facility_ID	Concatenated field including patient class	Concatenate the following values, with no separator: • C_Facility_ID (Site ID + Facility ID) • C_Patient_Class (E, I, O)

Example: Facility IDs

Archive Processed Table						
Sending Facility ID	**Sending Source**	**Treating Facility ID**	**C_Facility_ID**	**C_Facility_ID _Source**	**C_Processed_Facility_ID**	
1111	MSH4-2	1111	00011111	EVN7	00011111E	
1112	MSH4-1	.	00011111	MSH4	00011111E	
2244	MSH4-2	9999	00019999	EVN7	00019999E	
EXCEPTIONS Table						
.	.	7777	00017777	EVN7	00017777E	\<Error code\>

Operational Crosswalk (MFT + Crosswalk)	
Incoming Facility ID	**Outgoing Facility ID**
1111	00011111
1112	00011111
9999	00019999

Assumptions:
- Site ID = 0001
- Calculated Patient Class = E

Error code for exceptions table: Sending and Treating Facility IDs from message not present or active in MFT

Patient ID

A processed Patient ID will select the first non-null patient ID value to define a unique patient ID (C_Unique_Patient_ID). Inputs into this selection are stored separately*, as well.

Element	Processing	Considerations
Medical Record Number (PID-3)	Patient identifier with type "MR" from PID-3	Fairly Reliable if Present - if the Medical Record Number is received in the A01 message, it will also come in any A08 messages.
Patient ID (PID-2.1) [Legacy]*	Patient identifier from PID-2, non-repeating field	Should be consistent if present – this field is non-repeating
First Patient ID (PID-3)	First non-null value from PID-3	This repeating field may not have the same IDs in the same order with each message update
Patient Account Number (PID-18)	Account number from PID-18	May not contain a traditional patient ID but may serve as one for analysis
Visit Number (PV1-19)	Visit number from PV1-19	May not contain a traditional patient ID but may serve as one for analysis

Example: Patient ID

Incoming Message

Segment	Value
Medical Record Number (PID-3)*	12345
Patient ID (PID-2.1) [Legacy]	NULL
First Patient ID (PID-3)*	112233
Patient Account Number (PID-18)*	NULL
Visit Number / Visit ID (PV1-19)*	5689

* Stored as separate columns in Processed Table

BioSense Stage 1   BioSense Platform Processed Table

Unique Patient ID – Legacy Processing	Unique Patient ID – New Processing (proposed) C_Unique_Patient_ID	Unique Patient ID Source – New Processing (proposed) C_Unique_Patient_ID_Source
112233	**12345**	**MedRecNo (PID-3)**

Calculated Visit Date

- Return the earliest date/time from the following segments:
  - Admit Date/Time (PV1-44)
  - Discharge Date/Time (PV1-45)
  - Procedure Date/Time (PR1-5)
  - Patient Death Date/Time (PID-29)
  - Recorded Date/Time of Message (EVN-7.2)
  - Date/Time of Message (MSH-7)

  Processed Table will store:
  - C_Visit_Date_Time (date & time)
  - C_Visit_Date (date only)

- Proposed processing change
  - Remove OBX-14 "Date Time of Observation" from algorithm
    - May store data from earlier visits

Example: Calculated Visit Date

Segment	Value
Admit Date/Time	**1/15/2016 15:00:00**
Discharge Date/Time	1/17/2016 7:00:00
Procedure Date/Time	1/16/2016 16:00:00
Patient Death Date/Time	NULL
Recorded Date/Time	1/16/2016 18:00:00
Date/Time of Message	1/16/2016 18:00:00
Calculated Visit Date/Time	**1/15/2016 15:00:00**
Calculated Visit Date	**1/15/2015**
Calculated Visit Date/Time Source	**Admit Date/Time**

Calculated Unique Visit ID

- Uniquely identifies a "Visit"

- Comparable to existing Unique_Visiting_ID in current processing

- Concatenate the following fields, do not use a separator
  - Calculated Facility ID (Site ID + Facility ID)
  - Calculated Unique Patient ID
  - Calculated Visit Date

Calculated Facility ID (C_Facility_ID)	Unique Patient ID (C_Unique_Patient_ID)	Calculated Visit Date (C_Visit_Date)	Calculated Unique Visit ID (C_Unique_Visit_ID)
00011111	123456	01012016	0001111112345601012016
12349999	9812	01152016	1234999981201152016

Processed Visit ID

- Alternate approach to uniquely identify a "Visit" based on visit setting (patient class)
- Processing: concatenate the following values with no separator
  - Processed Facility ID (Site ID + Facility ID + Patient Class)
  - Calculated Unique Patient ID
  - Calculated Visit Date

Processed Facility ID (C_Processed_Facility_ID)	Unique Patient ID (C_Unique_Patient_ID)	Calculated Visit Date (C_Visit_Date)	Processed Visit ID (C_Processed_Visit_ID)
00011111E	123456	01012016	00011111E1234561012016
12349999O	9812	01152016	12349999O981201152016

Calculated Unique Visit ID (C_Unique_Visit_ID)	Processed Visit ID (C_Processed_Visit_ID)
000111111234561012016	00011111E1234561012016

Recap: BioSense Platform Data Flow Goals

- **Enable public health officials to conduct syndromic surveillance more systematically and collaboratively**

- **Provide access to data delivered to the BioSense Platform**
  - HL7 messages
  - ASCII converted to HL7 messages

- **Provide access to stable structured databases reflecting processed HL7 messages**

- **Prime the data with various calculated fields to allow for flexibility in analysis approaches regardless of the analytic tool being used**

- **Make the data complete, useful, and actionable**

NEXT STEPS

Mike Coletta, CDC

Processing Questions

- **What fields should be included in the algorithm to define a "Calculated Chief Complaint"?**

- **Discuss and vet processing for**
  - Calculated Patient Class
  - Calculated Facility ID
  - Processed Facility ID
  - Calculated Unique Patient ID
  - Calculated Unique Visit ID
  - Processed Visit ID

- **Other questions or concerns?**

Next Steps

- **Call for 6-9 volunteers to refine BioSense Platform Archive Data Element Requirements through small working sessions**
- **Schedule next webinar to discuss data elements to move into the ESSENCE application**

How do I volunteer?

Contact Mike Coletta by
2/15/16

mcoletta@cdc.gov

We appreciate your involvement.
Please feel free to contact

Michael A. Coletta, MPH
National Syndromic Surveillance Program Manager
CDC/CSELS/DHIS
mcoletta@cdc.gov

For more information please contact Centers for Disease Control and Prevention

1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
Visit: www.cdc.gov | Contact CDC at: 1-800-CDC-INFO or www.cdc.gov/info

Center for Surveillance, Epidemiology, and Laboratory Services
Division of Health Informatics and Surveillance