I2b2 ETL Breakout

JUNE 28, 2018 9:30 AM ROOM TBD

Agenda

I2b2	10 mins	Michael Mendis
TranSMART	10 mins	Peter Rice
Best Practices	15 mins	Lori, Michael
Tools	15 mins	Victoria, Michael, Victor
Open Discussion	30 mins	

I2b2 - Populating the Patient Mapping table

The **patient_mapping table** maps the *i2b2* ***patient_num*** to an encrypted number, ***patient_ide,*** from the *source_system* (the 'e' in ide is for 'encrypted').

The ***patient_ide_source*** contains the name of the source system.
The ***patient_ide_status*** gives the status of the patient number in the source system.

For example, if it is *Active, Inactive, Deleted,* or *Merged*.

patient_ide	patient_ide_source	patient_num	patient_ide_status
1000000	EMPI	528	A
123	MGH	528	A
777	BWH	528	A
528	HIVE	528	A
SDFDHSJKAHDSJDAS	MGH_E	528	A

I2b2 - Creating views for i2b2 projects

1) Create project in the admin tool
2) Add entry to CRC_DB_LOOKUP, ONT_DB_LOOKUP and WORK_DB_LOOKUP with the c_db_datasource having the same datasource as the main project, but the C_DB_FULLSCHEMA pointing to the new view
3) On the database create a new schema and populate all the QT tables
4) Create a database view to the main database for the OBSERVATiON_FACT, VISIT_DIMENSION, CONCEPT_DIMENSION, and possibly the PATIENT_DIMENSION

I2b2 – ETL decisions at BIDMC

Raw local codes v. standard ontology

- BIDMC maps to standard ontologies as part of ETL – This causes some information to be lost, but the i2b2 ontology is easier to manage, and it is less work to plugin to federated/SHRINE networks.

Full v. partial data updates

- BIDMC does a full data refresh with each update – Updates are done once a month, and the ETL takes less than one day. ETL processing occurs in staging tables so that i2b2 can remain running at all times.

Patient mapping table v. separate database to store MRNs

- BIDMC stores the mapping between MRNs and i2b2 patient_nums outside of i2b2 so that the i2b2 database by itself can be a limited dataset. This simplifies IRB protocols.

Separate databases v. views for i2b2 derived projects

- BIDMC creates separate databases for IRB reasons, even though this requires much more storage.

Single v. multiple fact tables (new i2b2 option)

- BIDMC uses a single fact table. With ~350M rows, a single table is small enough for good performance.

TranSMART - Peter

ETL Data Types

- Clinical
- High-dimensional
  - mRNA expression
  - RNAseq expression
  - Mass-spec proteomics
  - Metabolomics
  - Copy number variation / aCGH
  - Genomic variation (SNP…)
  - miRNA (qPCR, RNAseq)
- GWAS analysis

ETL Tools

- Kettle
- tMDataLoader
- Transmart-batch
- …other

TranSMART - Kettle

Kettle scripts (Pentaho data-integration)

- Original tranSMART ETL method
- Scripts to manipulate data
  - Pivot input files
  - Load to staging tables
  - Invoke stored procedures
  - Try to catch errors
- Integrated in other tools
  - Transmart-data make targets
  - ICE tool GUI

Data sources

- 200+ studies (GEO etc.) on library.transmartfoundation.org
- Loading scripts for each study
  - Clinical
  - Ref_annotation
  - Expression
  - …

TranSMART - tMDataLoader

tMDataLoader

- Developed by Clarivate (formerly Thomson Reuters)
- Validates input files
- Adds new options for clinical data
- Run loads all data types for all studies

Issues

- Uses copy of stored procedures
- Supporting alternative versions of tranSMART (16.3, 17.1 server)
- Requires changes to tranSMART schema
- Renames input data files/directories
  - Rename back to re-run

TranSMART – transmart-batch

Transmart-batch

- Developed by the Hyve (Netherlands)
- Replaces stored procedures
- Avoids staging tables
- Designed for performance
- Repeat previous run

Issues

- Extends tranSMART schema
  - Included in postgreSQL schema
  - Need to extend for Oracle
- Active development
  - Check for changes in new versions

Best Practices – Ontologies

Mapping for Diagnose and Procdures (using in standard format ICD

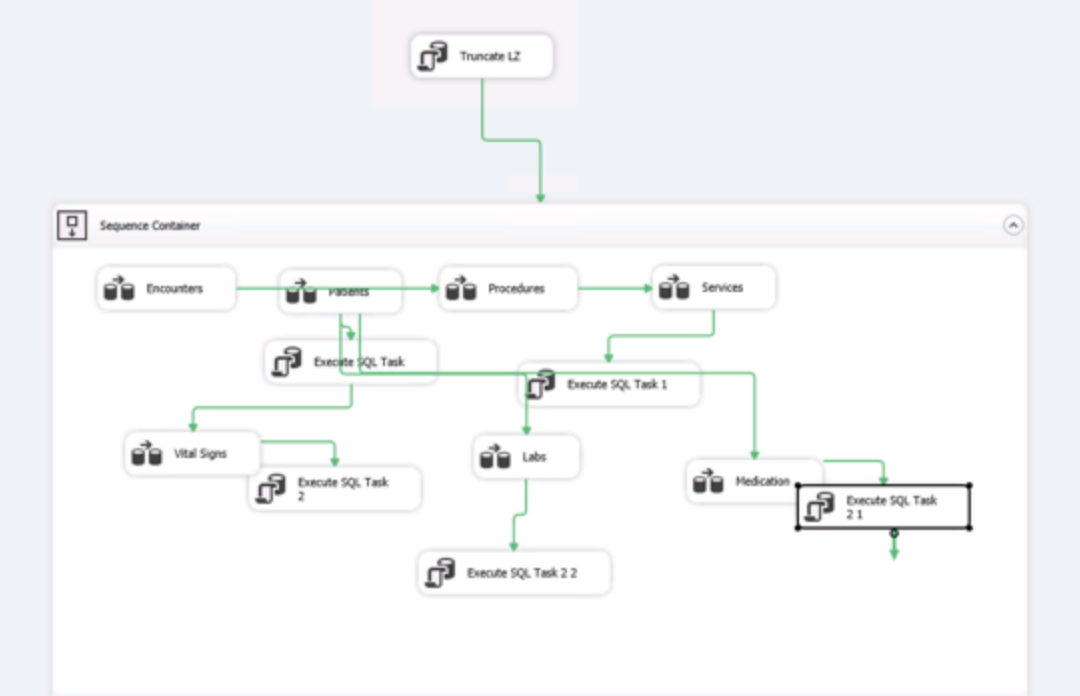
Mapping for Meds and Labs which are not in standard format (mike talks about java side)

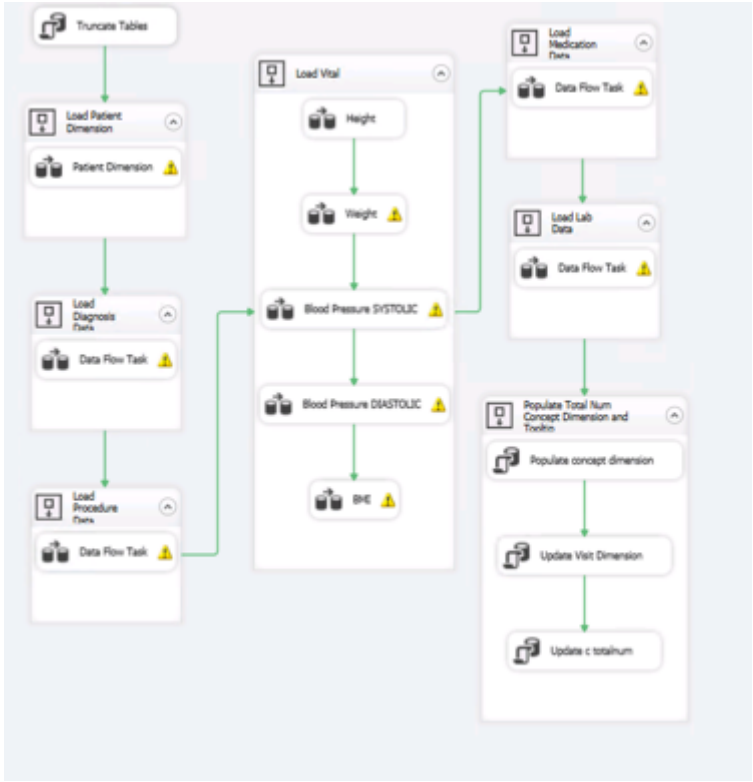
Ontologies that can be used Ontology Group (Lori)

Data Load ETL

Best Practices - Loading & Working Zone

Load data as a direct copy from the source EHR system, convert to a standardize format, and than load that data into i2b2 or any other datawarehouse.



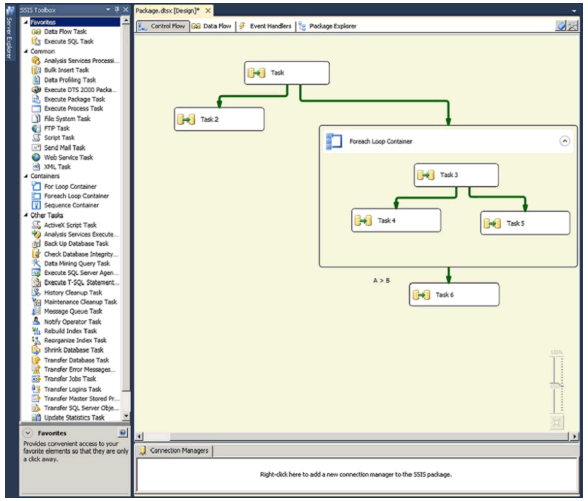


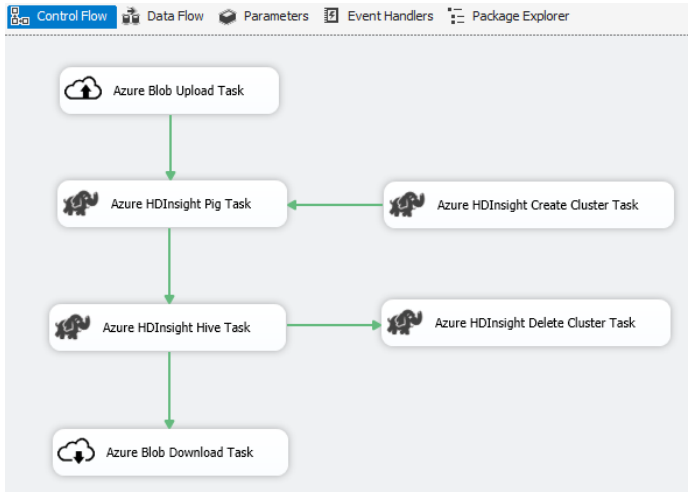
Tools – SSIS (SQL Server Integration Services)

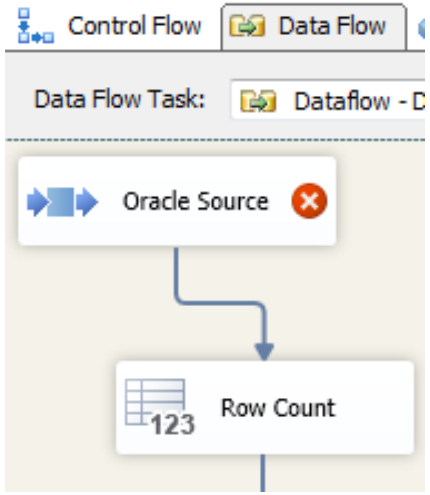
Microsoft Integration Services is a platform for building enterprise-level data integration and data transformations solutions.

SQL Server Data Tools is a modern development tool that you can download for free to build SQL Server relational databases, Azure SQL databases, Integration Services packages, Analysis Services data models, and Reporting Services reports.

https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-2017







Tools - Hive

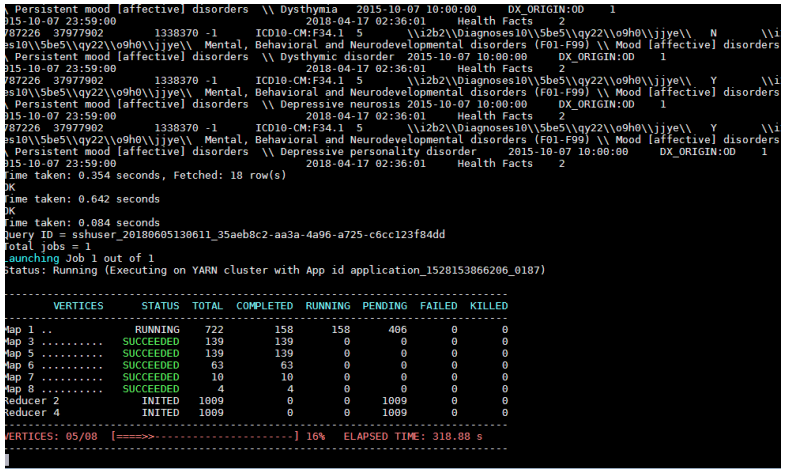
Hive is a Built in Database tool provided in Hadoop environments (beeline is the new CLI).

Hive can feel intimidating, but it is simply a matter of planning out your steps.

We use as our base, a data dump from our EHR I2B2, which we transform and supplement with additional data from other sources and then populate into our internal I2B2 instance, which is a later version with multiple fact tables.

This ingestion process has two main issues to over come, first is the use of || as the separator, coming from supplied .dat files

Second and the trickiest part of using Hive is dealing with the “\ “ character, which is a reserved character in Hive, so you need to escape it.



Tools - Hive

```
DROP VIEW IF EXISTS lci.charnc_ont_i2b2_loinc_s3;

CREATE VIEW lci.charnc_ont_i2b2_loinc_s3 AS
SELECT
conc.c_hlevel,
conc.C_FULLNAME,
conc.c_name,
conc.C_SYNONYM_CD,
conc.C_VISUALATTRIBUTES,
conc.C_TOTALNUM,
conc.C_BASECODE,
conc.C_METADATAXML,
```

RESULTS

	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b2_loinc_s3.c_factta...	charnc_ont_i2b...	charnc_ont_i2b...	charnc_ont_i2b...	charnc
1	7	\\i2b2\\Laborat...	Codfish IgE Qn ...	N	LH	0	LOINC:6082-2	<?xml version=...	loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
2	7	\\i2b2\\Laborat...	Codfish IgE-Rto...	N	LH	0	LOINC:39081-5	<?xml version=...	loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
3	7	\\i2b2\\Laborat...	Codfish IgG RA...	N	LH	0	LOINC:21213-4		loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
4	7	\\i2b2\\Laborat...	Codfish IgG4-m...	N	LH	0	LOINC:56228-0	<?xml version=...	loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
5	6	\\i2b2\\Laborat...	Codfish (Gadus ...	N	FH	0	LOINC:LP62344...		loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
6	7	\\i2b2\\Laborat...	Codfish LR-mCn...	N	LH	0	LOINC:48224-0	<?xml version=...	loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
7	5	\\i2b2\\Laborat...	Crab (Cancer pa...	N	FH	0	LOINC:LP16960...		loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
8	1	\\i2b2\\Laborat...	Laboratory Test...	N	CA	1737944			loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
9	2	\\i2b2\\Laborat...	Allergy (LP3162...	N	FH	0	LOINC:LP31625...		loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE
10	5	\\i2b2\\Laborat...	Carp | Bld-Ser-P...	N	FH	0	LOINC:LP46763...		loinc_fact.concept_cd	concept_dimen...	concept_path	T	LIKE

```
conc.C_PATH,
conc.C_SYMBOL

FROM lci.charnc_ont_i2b2_loinc_s2 conc WHERE conc.c_name <>'C_NAME';

SELECT * FROM lci.charnc_ont_i2b2_loinc_s3 LIMIT 10;
```

Tools - Hive

Once you have completed the transformation in hive to insert the data back into the working database tables,

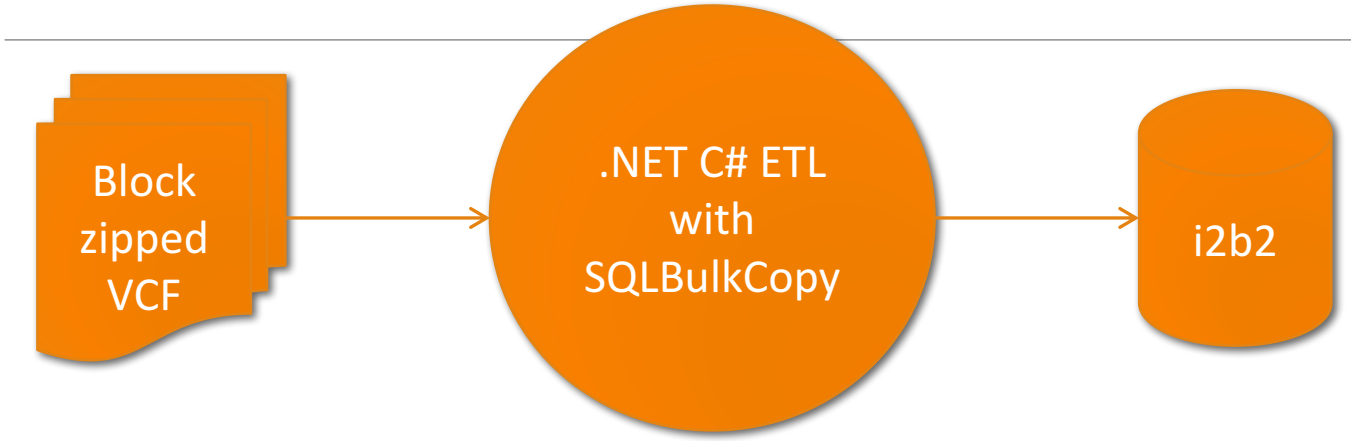
you can use tools like **Sqoop** (works with Oracle, Postgres and MSSQL Server)

or specific to the database tools like **bcp** to move the data into MSQL Server

JCNRUNBCPI2B2FILES.ps1*

```
clear

$SQLCMDPath = "C:\Program Files\Microsoft SQL Server\110\Tools\Binn\sqlcmd.exe"

$BCPPath = "C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\130\Tools\Binn\bcp.exe"

$BCPDelimiter = "||"

$ImportFilePath = "F:\Incoming\CNRI2B2\dat_cerner_files\"

$SqlServer = "
$SqlDatabase = "i2b2chsmetadata"
$SqlDatabaseB = "i2b2chsdata"
$SqlUser =
$SqlPassword =
$files = @("charnc_ont_I2B2_CPT4.dat","charnc_ont_I2B2_LOINC.dat","charnc_ont_I2B2_NCD.dat","charnc_ont_I2B2_
$tables = @("I2B2_CPT4","I2B2_LOINC","I2B2_NCD","I2B2_VISIT")

#run through the ontology tables that may have changed

for($i=0; $i -lt $tables.Length; $i++){

& $SQLCMDPath -S $SqlServer -d $SqlDatabase -U $SqlUser -P $SqlPassword -Q "truncate table dbo.$tables[$i]"

$Fullname= join-path -path $ImportFilePath -childpath $files[$i]
Write-Host $Fullname

& $BCPPath $tables[$i] in $Fullname  -S $SqlServer -d $SqlDatabase -U $SqlUser -P $SqlPassword -c -t "$BCPDel

}

#now do the OBS FACT TABLES
```

Tools – Genomics - Victor



#CHROM	POS	ID	REF	ALT	...	INFO	...	SUBJECT_1
1	752566	rs3094315	G	A	...	RSID=rs3094315;VariantEffect=FAM87B|NR_103536.1:n.-185G>A|p.=|upstream	...	1/1

CONCEPT_CD	INSTANCE_NUM	VALTYPE_CD	OBSERVATION_BLOB
SO:0001483	1	B	rs3094315,G_to_A,FAM87B,homozygous,upstream

CONCEPT_CD

- Two concepts with codes from Sequence Ontology: SNP (SO:0001483) or indel (SO:1000032)

INSTANCE_NUM

- The set of all SNPs for each patient will all have the same encounter number and date
- The concept codes will be the same for all SNPs (SO:0001483) and for all indels (SO:1000032).
- The set of all SNP facts will be enumerated in the instance_num field to make the primary key unique, as will the set of all indels

VALTYPE_CD

- always equal “B” to indicate that data are stored in the observation_blob field and to trigger the full text search already existing in the i2b2 environment

CONCEPT_CD	INSTANCE_NUM	VALTYPE_CD	—LARGESTRING search of OBSERVATION_BLOB
			OBSERVATION_BLOB
SO:0001483	1	B	rs3094315,G_to_A,FAM87B,homozygous,upstream
SO:0001483	2	B	rs3131972,A_to_G,FAM87B,homozygous,upstream
SO:0001483	3	B	rs61770172,C_to_G,FAM87B,homozygous,exon
SO:0001483	4	B	rs3115860,C_to_A,FAM87B,homozygous,exon
SO:0001483	5	B	rs12567639,G_to_A,FAM87B,homozygous,downstream
SO:0001483	6	B	rs377214516,C_to_T,LINC01128,homozygous,upstream
SO:0001483	7	B	rs540936498,C_to_T,LINC00115,homozygous,exon

OBSERVATION_BLOB

<RSID|“missing_rsid”>,<REF_TO_ALT>,<GENE_SYMBOL|“missing_gene”>,<ZYGOSITY|“missing_zygosity”>,<CONSEQUENCE|“missing_consequence”>

Tools – FLAT FHIR

- FHIR (Fast Healthcare Interoperability Resources) (http://hl7.org/fhir/overview.html)
- The generally available version right now is V3
- Bulk Updates (Flat FHIR) is a new API proposed as part of R4 Balllot #1 (planned for 2018)
- Allows for a Bulk request for **Patient Everything** (single patient or a group) (all supported elements for all time), (**http://hl7.org/fhir/2018May/group-operation-everything.html)**
- Usually completed Asynchronously – the output format currently proposed is ND-Json (New line delimited JSON)
- Transformation would be required to get it into the STAR I2B2 schema, and you would have to develop a process for ingestion, but there is no reason, once this format becomes more generally available that it could not be used as a data source like any other.

Tools – FLAT FHIR- Links

http://docs.smarthealthit.org/flat-fhir/-

http://www.healthintersections.com.au/?p=2689

http://hl7.org/fhir/2018May/group-operation-everything.html

http://hl7.org/fhir/2018May/operation-patient-everything.html

http://hl7.org/fhir/2018May/async.html

http://hl7.org/fhir/2018May/formats.html#bulk

Open Discussion

Best Practices – Refreshes

Other Tools people have used

IRB

Other ways of doing ETL

Look at BI and have disuccsion on the 5 items