Course 2 Module 5
Programming Assignment

Assignment is to ETL MIMIC data into the OMOP CONDITION_OCCURRENCE table

Detailed instructions with Slide Notes

Assignment is to ETL MIMIC data into the OMOP CONDITION_OCCURRENCE table

ETL Steps

1. Understand source/target data models
2. Profile source tables
3. Create ETL mappings
4. Write transformation code
5. Execute transformation
6. Perform data quality assessment
7. Package documentation

Step 1: Understand source/target data models

CONDITION_OCCURRENCE is the TARGET OMOP table.

Read the OMOP documentation about the type of data stored in CONDITION_OCCURRENCE and for three fields below that are in that table:

- **person_id**
- **visit_occurrence_id**
- **condition_source_value**

Table Details: condition_occurrence

Schema | Details | Preview

condition_occurrence_id	FLOAT	NULLABLE	int64
person_id	FLOAT	NULLABLE	int64
condition_concept_id	FLOAT	NULLABLE	int64
condition_start_date	STRING	NULLABLE	parse_date()
condition_start_datetime	STRING	NULLABLE	parse_datetime()
condition_end_date	STRING	NULLABLE	parse_date()
condition_end_datetime	STRING	NULLABLE	parse_datetime()
condition_type_concept_id	FLOAT	NULLABLE	int64
stop_reason	STRING	NULLABLE	Describe this field...
provider_id	FLOAT	NULLABLE	int64
visit_occurrence_id	FLOAT	NULLABLE	int64
visit_detail_id	FLOAT	NULLABLE	int64
condition_source_value	STRING	NULLABLE	Describe this field...
condition_source_concept_id	FLOAT	NULLABLE	int64
condition_status_source_value	STRING	NULLABLE	Describe this field...
condition_status_concept_id	FLOAT	NULLABLE	int64

Step 2: Profile source table or tables

Using the White Rabbit profiling data from the 100 patient MIMIC database provided in the Assessment to comment on the distribution of the SUBJECT_ID field from one of the MIMIC tables selected in Step 1

- MIMIC TableName DIAGNOSES_ICD
  - There are no missing values in subject_id, hadm_id, or ICD9_code
  - Number of admissions for subject_id ranges from 5-266, but the list is truncated

Step 3: Create ETL mappings

Table Details: DIAGNOSES_ICD

Schema | Details | Preview

ROW_ID	INTEGER	NULLABLE	Describe tl
SUBJECT_ID	INTEGER	NULLABLE	Describe tl
HADM_ID	INTEGER	NULLABLE	Describe tl
SEQ_NUM	INTEGER	NULLABLE	Describe tl
ICD9_CODE	STRING	NULLABLE	Describe tl

Table Details: condition_occurrence

Schema | Details | Preview

condition_occurrence_id	FLOAT	NULLABLE	int64
person_id	FLOAT	NULLABLE	int64
condition_concept_id	FLOAT	NULLABLE	int64
condition_start_date	STRING	NULLABLE	parse_date()
condition_start_datetime	STRING	NULLABLE	parse_datetime()
condition_end_date	STRING	NULLABLE	parse_date()
condition_end_datetime	STRING	NULLABLE	parse_datetime()
condition_type_concept_id	FLOAT	NULLABLE	int64
stop_reason	STRING	NULLABLE	Describe this field...
provider_id	FLOAT	NULLABLE	int64
visit_occurrence_id	FLOAT	NULLABLE	int64
visit_detail_id	FLOAT	NULLABLE	int64
condition_source_value	STRING	NULLABLE	Describe this field...
condition_source_concept_id	FLOAT	NULLABLE	int64
condition_status_source_value	STRING	NULLABLE	Describe this field...
condition_status_concept_id	FLOAT	NULLABLE	int64

All codes are from the DIAGNOSES_ICD table

I choose the Subject_ID to map to the person_ID. Both are unique identifiers for an individual patient.

I chose the HADM_ID to correspond to the visit_occurrence_id. Both are unique identifiers for when a condition diagnosis is made.

I chose the ICD9_CODE from DIAGNOSES_ICD to correspond to the condition_source_value. Both represent the ICD9 code of the condition.

Step 3: 4th ETL mapping

Table Details: ADMISSIONS

Schema | Details | Preview

ROW_ID	INTEGER
SUBJECT_ID	INTEGER
HADM_ID	INTEGER
ADMITTIME	DATETIME
DISCHTIME	DATETIME
DEATHTIME	DATETIME
ADMISSION_TYPE	STRING
ADMISSION_LOCATION	STRING
DISCHARGE_LOCATION	STRING
INSURANCE	STRING
LANGUAGE	STRING
RELIGION	STRING
MARITAL_STATUS	STRING
ETHNICITY	STRING
EDREGTIME	DATETIME
EDOUTTIME	DATETIME
DIAGNOSIS	STRING
HOSPITAL_EXPIRE_FLAG	INTEGER
HAS_CHARTEVENTS_DATA	INTEGER

from the Admissions table

I choose the ADMITTIME to map to CONDITION_START_DATETIME. Both track the beginning of the medical visit (OMOP)/ER visit (MIME).

Table Details: condition_occurrence

Schema | Details | Preview

condition_occurrence_id	FLOAT	NULLABLE	int64
person_id	FLOAT	NULLABLE	int64
condition_concept_id	FLOAT	NULLABLE	int64
condition_start_date	STRING	NULLABLE	parse_date()
condition_start_datetime	STRING	NULLABLE	parse_datetime()
condition_end_date	STRING	NULLABLE	parse_date()
condition_end_datetime	STRING	NULLABLE	parse_datetime()
condition_type_concept_id	FLOAT	NULLABLE	int64
stop_reason	STRING	NULLABLE	Describe this field...
provider_id	FLOAT	NULLABLE	int64
visit_occurrence_id	FLOAT	NULLABLE	int64
visit_detail_id	FLOAT	NULLABLE	int64
condition_source_value	STRING	NULLABLE	Describe this field...
condition_source_concept_id	FLOAT	NULLABLE	int64
condition_status_source_value	STRING	NULLABLE	Describe this field...
condition_status_concept_id	FLOAT	NULLABLE	int64

Step 4: Write transformation code

Paste the SQL statements that transform data from one or more MIMIC tables into the three OMOP CONDITION_OCCURRENCE fields (patient-id, visit_occurrence_id, condition_source_value) into the Coursera Submission Site

Step 5: Execute transformation code

Execute the ETL code from Step 4 but do not submit the output table.

Use the output table for Step 6.

There is no submission for this Step.

Step 6: Perform data quality assessment

Define, implement, execute one or more data quality measures. Submit final DQ measure and an explanation why you created your measure(s).

Row	MaxVisits	MinVisits	AvgVisits
1	266	3	17.6

Data quality measures implemented: check for missing values and for each patient and count admissions per patient. Across all patients what was minimum (3), maximum (266), average (17.6) and median (13) number of admissions. These data quality measures were implemented to see if data appeared to have been extracted correctly. For ICU patients it is not surprising that each patient visited the hospital more than once. This confirms the population is drawn from ICU admittance. The average admission value is skewed toward the outlier max admission value. In this case the median should be used instead if a representative patient admission number needs to be calculated.

Step 7: Package documentation

- Congratulations! The materials in the previous slides constitute a complete ETL package.

There is no submission for this Step.