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**

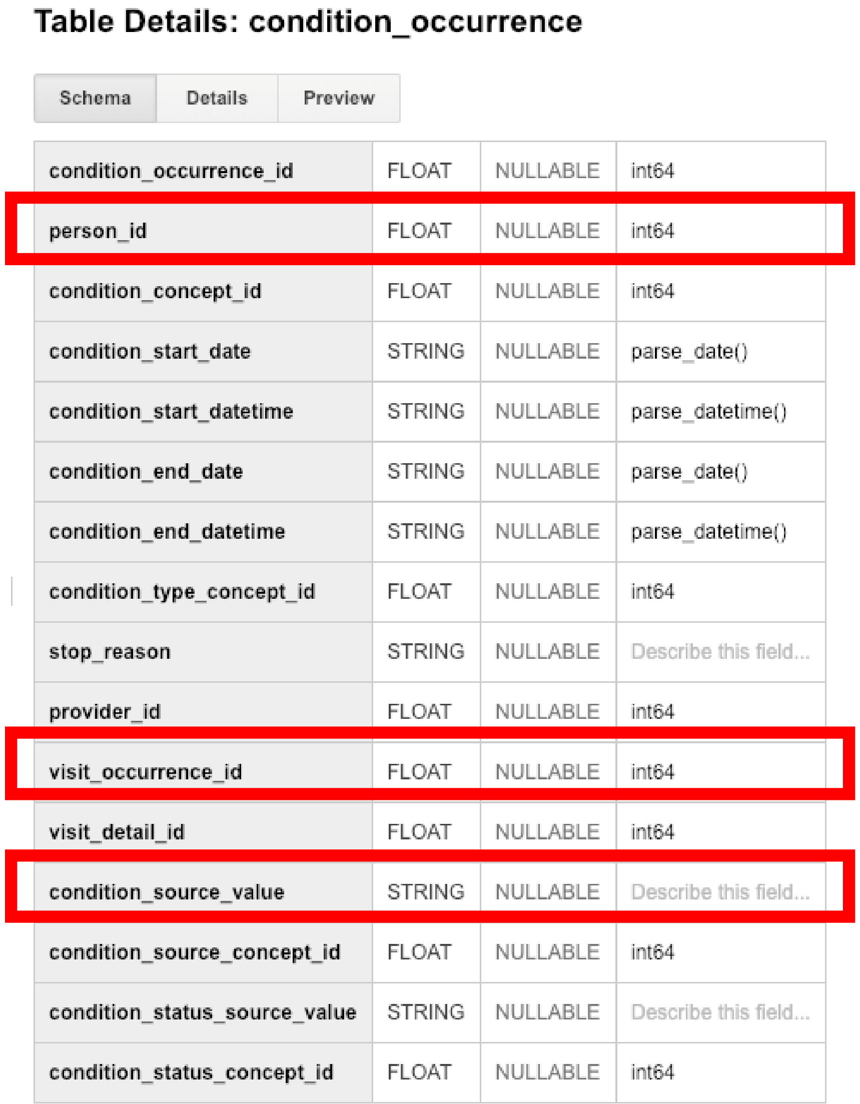


Table Details: condition_occurrence

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 1: Understand source/target data models

CONDITION_OCCURRENCE is the TARGET OMOP table.

Select one or more MIMIC tables from the table screen shots on the next slides that you feel are most related to the three fields in CONDITION_OCCURRENCE.

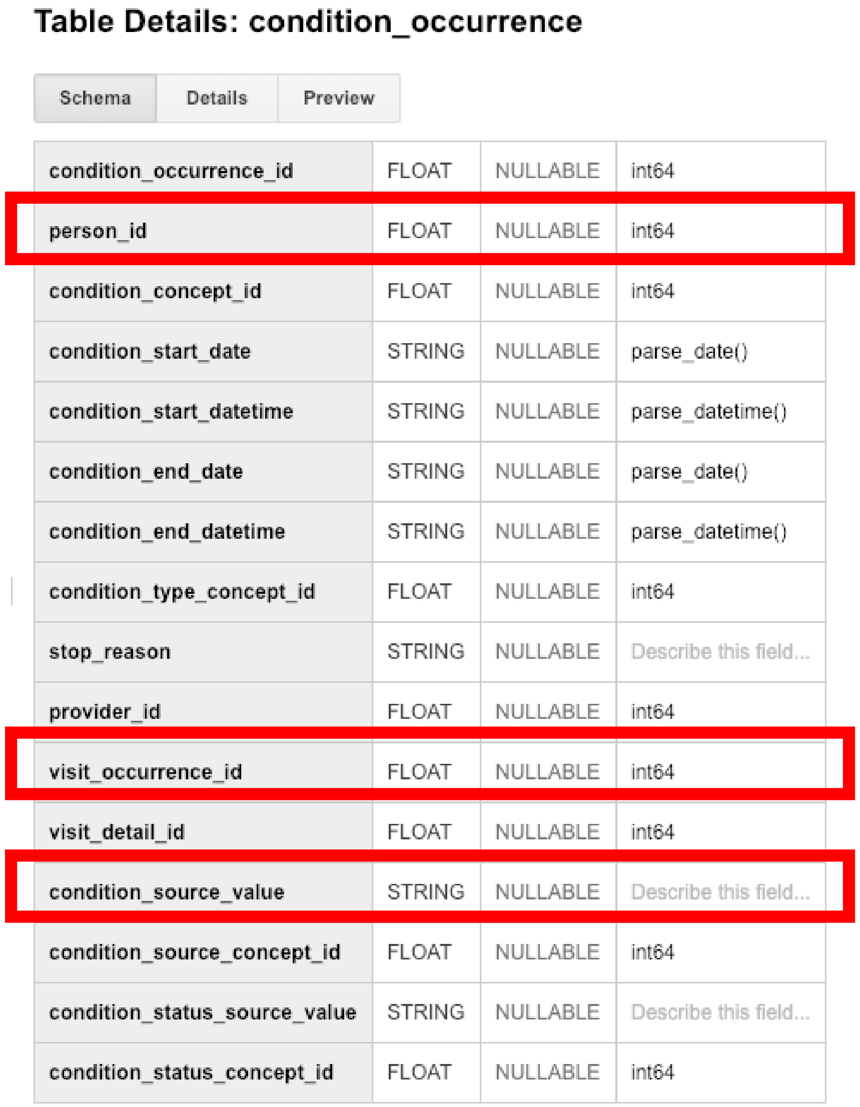


Table Details: condition_occurrence

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

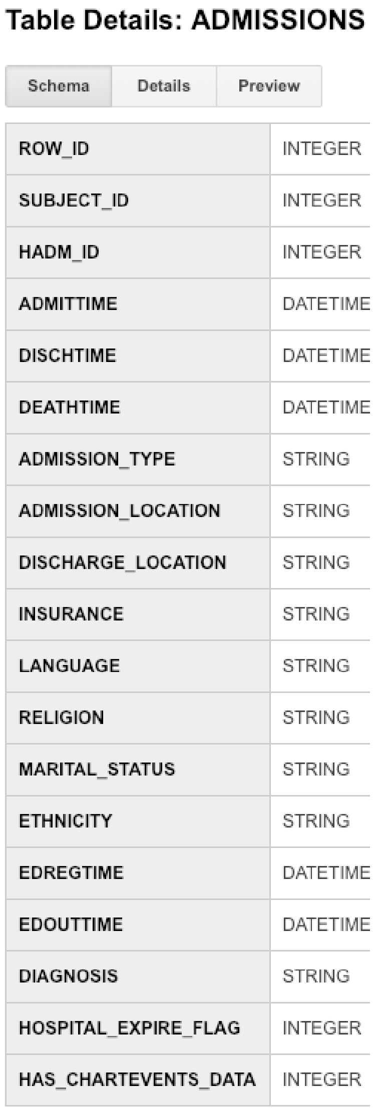


Table Details: ADMISSIONS

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

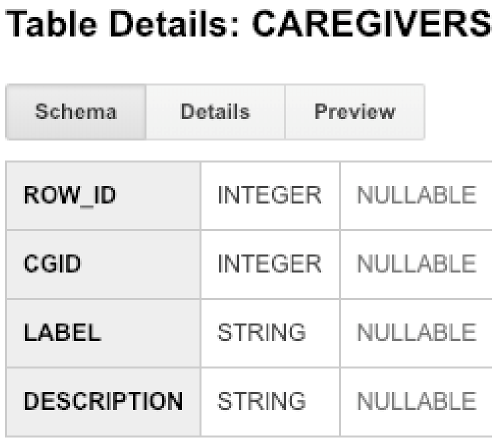


Table Details: CAREGIVERS

Schema Details Preview

ROW_ID	INTEGER	NULLABLE
CGID	INTEGER	NULLABLE
LABEL	STRING	NULLABLE
DESCRIPTION	STRING	NULLABLE

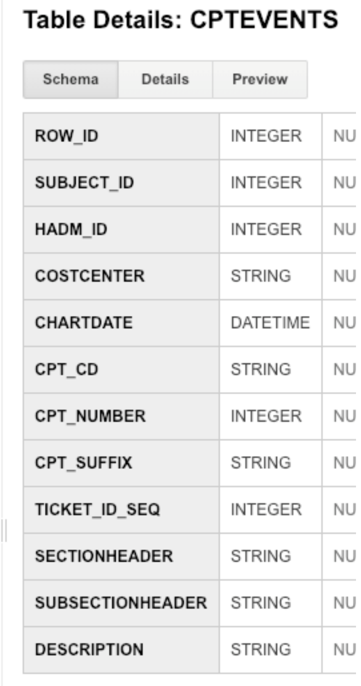


Table Details: CPTEVENTS

Schema Details Preview

ROW_ID	INTEGER	NU
SUBJECT_ID	INTEGER	NU
HADM_ID	INTEGER	NU
COSTCENTER	STRING	NU
CHARTDATE	DATETIME	NU
CPT_CD	STRING	NU
CPT_NUMBER	INTEGER	NU
CPT_SUFFIX	STRING	NU
TICKET_ID_SEQ	INTEGER	NU
SECTIONHEADER	STRING	NU
SUBSECTIONHEADER	STRING	NU
DESCRIPTION	STRING	NU

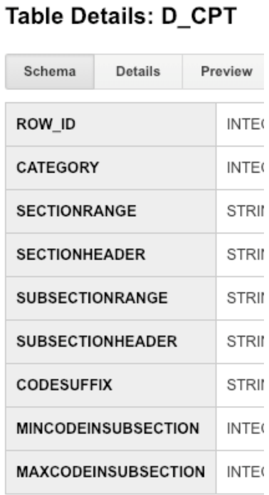


Table Details: D_CPT

Schema Details Preview

ROW_ID	INTE
CATEGORY	INTE
SECTIONRANGE	STRI
SECTIONHEADER	STRI
SUBSECTIONRANGE	STRI
SUBSECTIONHEADER	STRI
CODESUFFIX	STRI
MINCODEINSUBSECTION	INTE
MAXCODEINSUBSECTION	INTE

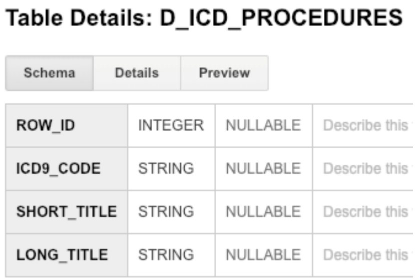


Table Details: D_ICD_PROCEDURES

Schema Details Preview

ROW_ID	INTEGER	NULLABLE	Describe this
ICD9_CODE	STRING	NULLABLE	Describe this
SHORT_TITLE	STRING	NULLABLE	Describe this
LONG_TITLE	STRING	NULLABLE	Describe this

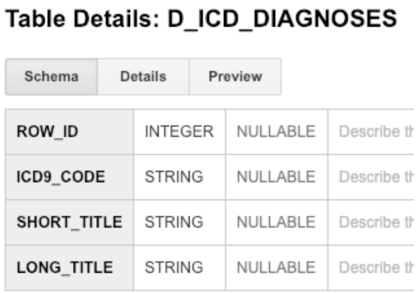


Table Details: D_ICD_DIAGNOSES

Schema Details Preview

ROW_ID	INTEGER	NULLABLE	Describe th
ICD9_CODE	STRING	NULLABLE	Describe th
SHORT_TITLE	STRING	NULLABLE	Describe th
LONG_TITLE	STRING	NULLABLE	Describe th

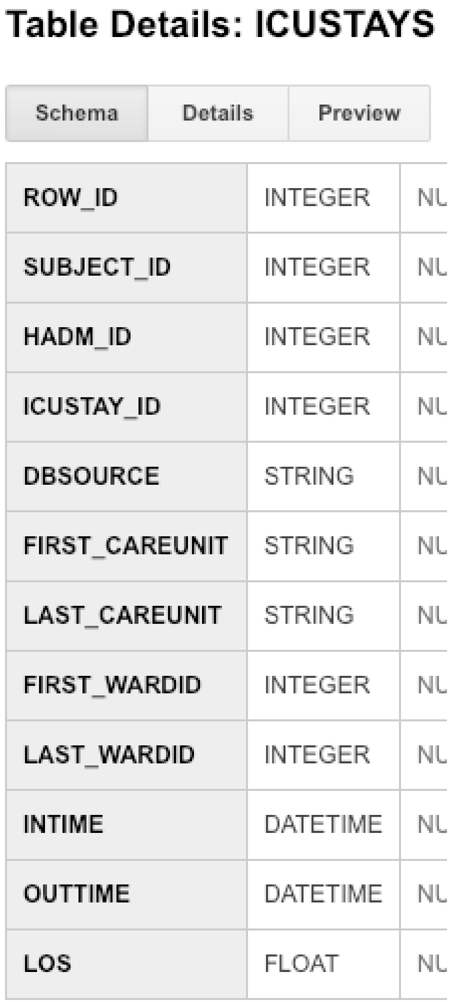


Table Details: ICUSTAYS

Schema Details Preview

ROW_ID	INTEGER	NU
SUBJECT_ID	INTEGER	NU
HADM_ID	INTEGER	NU
ICUSTAY_ID	INTEGER	NU
DBSOURCE	STRING	NU
FIRST_CAREUNIT	STRING	NU
LAST_CAREUNIT	STRING	NU
FIRST_WARDID	INTEGER	NU
LAST_WARDID	INTEGER	NU
INTIME	DATETIME	NU
OUTTIME	DATETIME	NU
LOS	FLOAT	NU

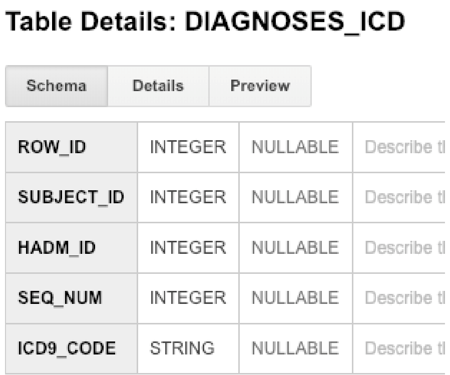


Table Details: DIAGNOSES_ICD

Schema Details Preview

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

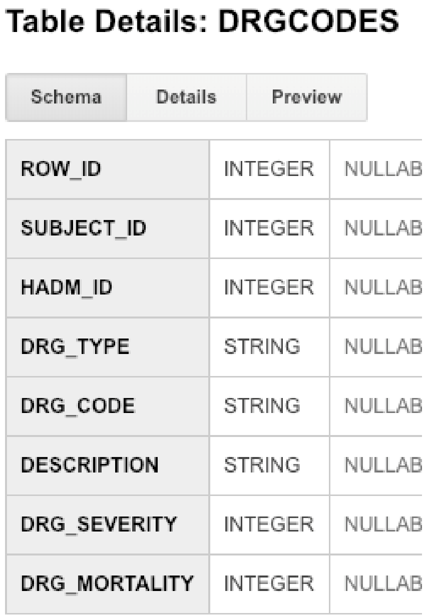


Table Details: DRGCODES

Schema Details Preview

ROW_ID	INTEGER	NULLAB
SUBJECT_ID	INTEGER	NULLAB
HADM_ID	INTEGER	NULLAB
DRG_TYPE	STRING	NULLAB
DRG_CODE	STRING	NULLAB
DESCRIPTION	STRING	NULLAB
DRG_SEVERITY	INTEGER	NULLAB
DRG_MORTALITY	INTEGER	NULLAB

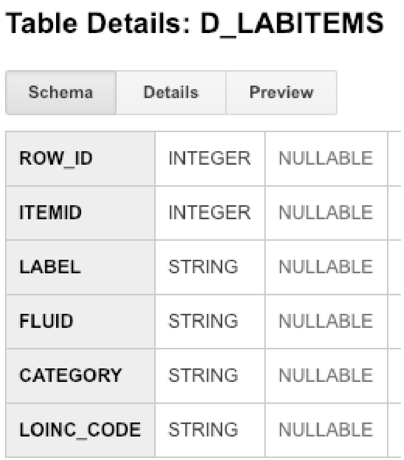


Table Details: D_LABITEMS

Schema Details Preview

ROW_ID	INTEGER	NULLABLE
ITEMID	INTEGER	NULLABLE
LABEL	STRING	NULLABLE
FLUID	STRING	NULLABLE
CATEGORY	STRING	NULLABLE
LOINC_CODE	STRING	NULLABLE

Use these screen captures (and next slide) to select one or more MIMIC tables that contain data for OMOP CONDITION_OCCURRENCE table

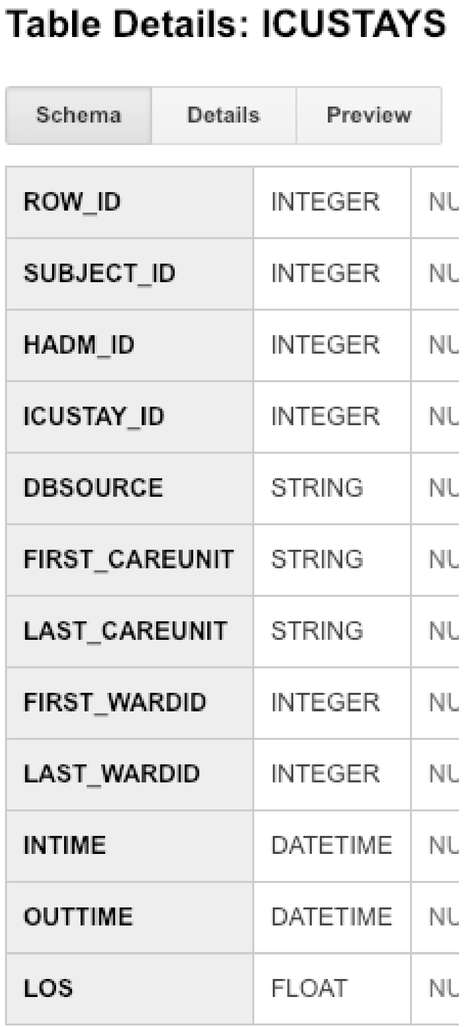


Table Details: ICUSTAYS

ROW_ID	INTEGER	NU
SUBJECT_ID	INTEGER	NU
HADM_ID	INTEGER	NU
ICUSTAY_ID	INTEGER	NU
DBSOURCE	STRING	NU
FIRST_CAREUNIT	STRING	NU
LAST_CAREUNIT	STRING	NU
FIRST_WARDID	INTEGER	NU
LAST_WARDID	INTEGER	NU
INTIME	DATETIME	NU
OUTTIME	DATETIME	NU
LOS	FLOAT	NU

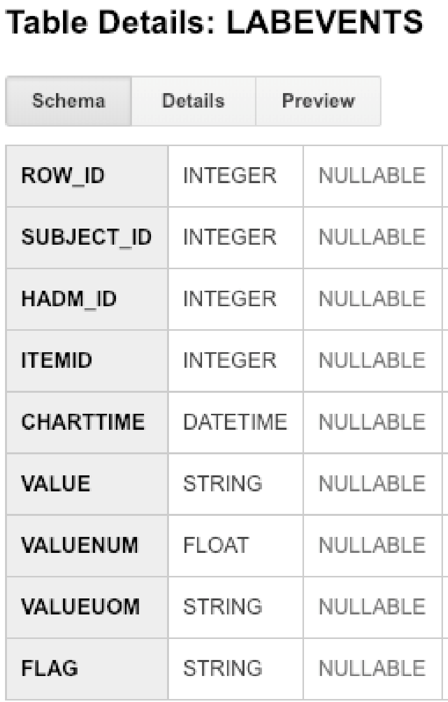


Table Details: LABEVENTS

ROW_ID	INTEGER	NULLABLE
SUBJECT_ID	INTEGER	NULLABLE
HADM_ID	INTEGER	NULLABLE
ITEMID	INTEGER	NULLABLE
CHARTTIME	DATETIME	NULLABLE
VALUE	STRING	NULLABLE
VALUENUM	FLOAT	NULLABLE
VALUEUOM	STRING	NULLABLE
FLAG	STRING	NULLABLE

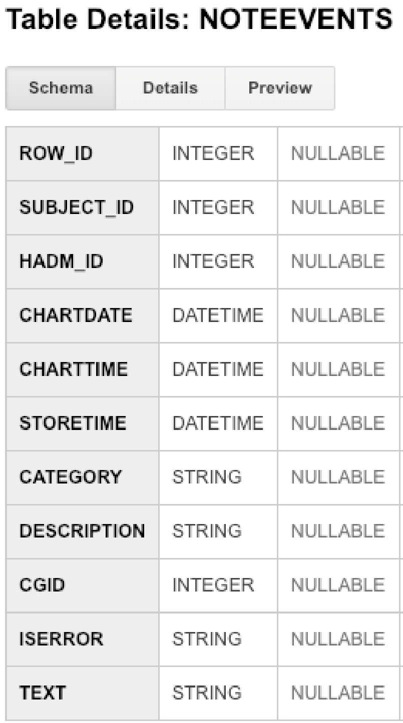


Table Details: NOTEEVENTS

ROW_ID	INTEGER	NULLABLE
SUBJECT_ID	INTEGER	NULLABLE
HADM_ID	INTEGER	NULLABLE
CHARTDATE	DATETIME	NULLABLE
CHARTTIME	DATETIME	NULLABLE
STORETIME	DATETIME	NULLABLE
CATEGORY	STRING	NULLABLE
DESCRIPTION	STRING	NULLABLE
CGID	INTEGER	NULLABLE
ISERROR	STRING	NULLABLE
TEXT	STRING	NULLABLE

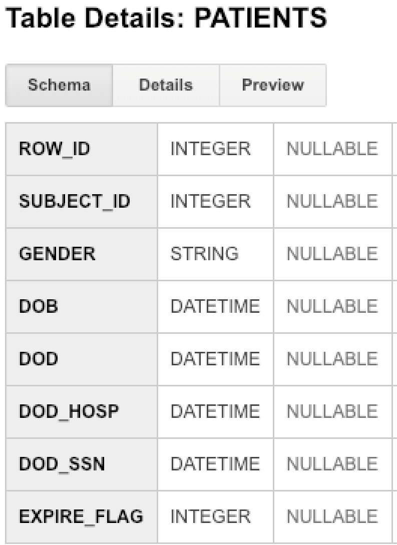


Table Details: PATIENTS

ROW_ID	INTEGER	NULLABLE
SUBJECT_ID	INTEGER	NULLABLE
GENDER	STRING	NULLABLE
DOB	DATETIME	NULLABLE
DOD	DATETIME	NULLABLE
DOD_HOSP	DATETIME	NULLABLE
DOD_SSN	DATETIME	NULLABLE
EXPIRE_FLAG	INTEGER	NULLABLE

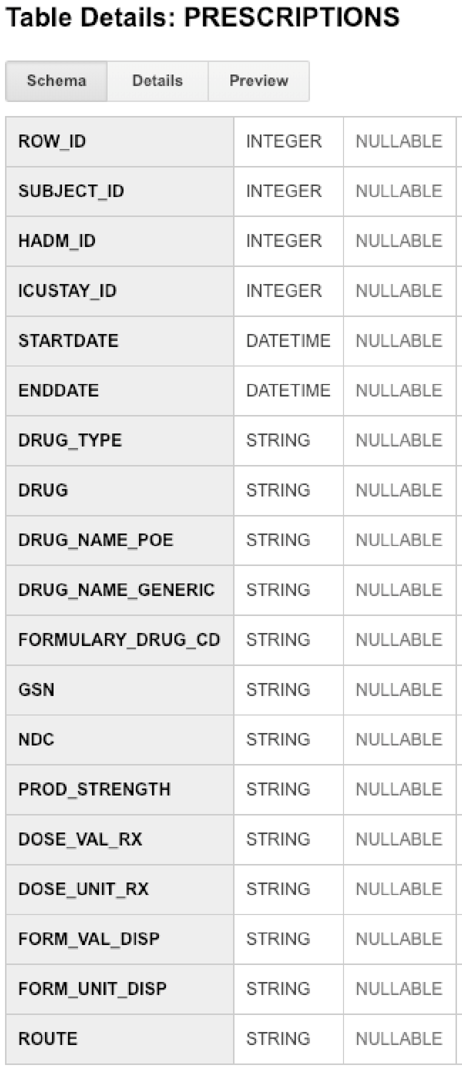


Table Details: PRESCRIPTIONS

ROW_ID	INTEGER	NULLABLE
SUBJECT_ID	INTEGER	NULLABLE
HADM_ID	INTEGER	NULLABLE
ICUSTAY_ID	INTEGER	NULLABLE
STARTDATE	DATETIME	NULLABLE
ENDDATE	DATETIME	NULLABLE
DRUG_TYPE	STRING	NULLABLE
DRUG	STRING	NULLABLE
DRUG_NAME_POE	STRING	NULLABLE
DRUG_NAME_GENERIC	STRING	NULLABLE
FORMULARY_DRUG_CD	STRING	NULLABLE
GSN	STRING	NULLABLE
NDC	STRING	NULLABLE
PROD_STRENGTH	STRING	NULLABLE
DOSE_VAL_RX	STRING	NULLABLE
DOSE_UNIT_RX	STRING	NULLABLE
FORM_VAL_DISP	STRING	NULLABLE
FORM_UNIT_DISP	STRING	NULLABLE
ROUTE	STRING	NULLABLE

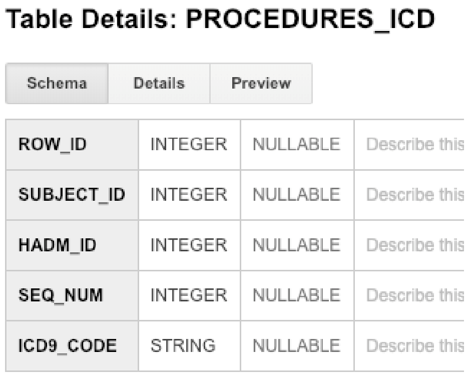


Table Details: PROCEDURES_ICD

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

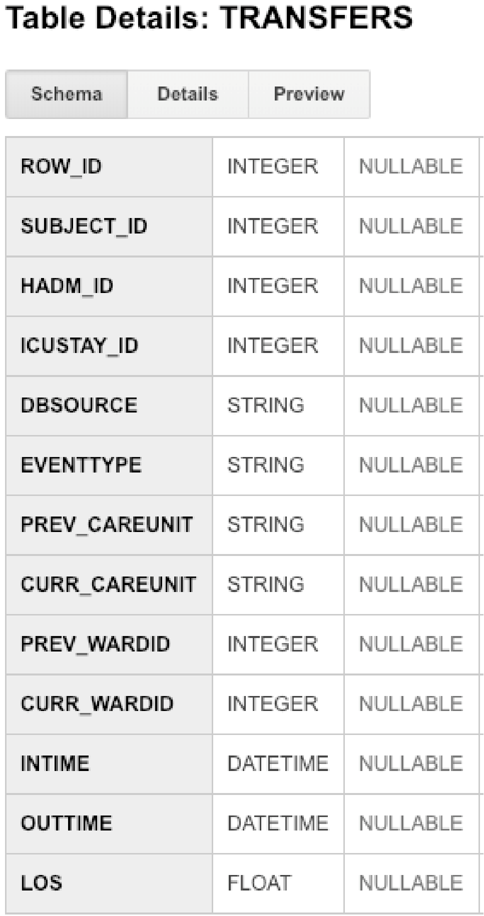
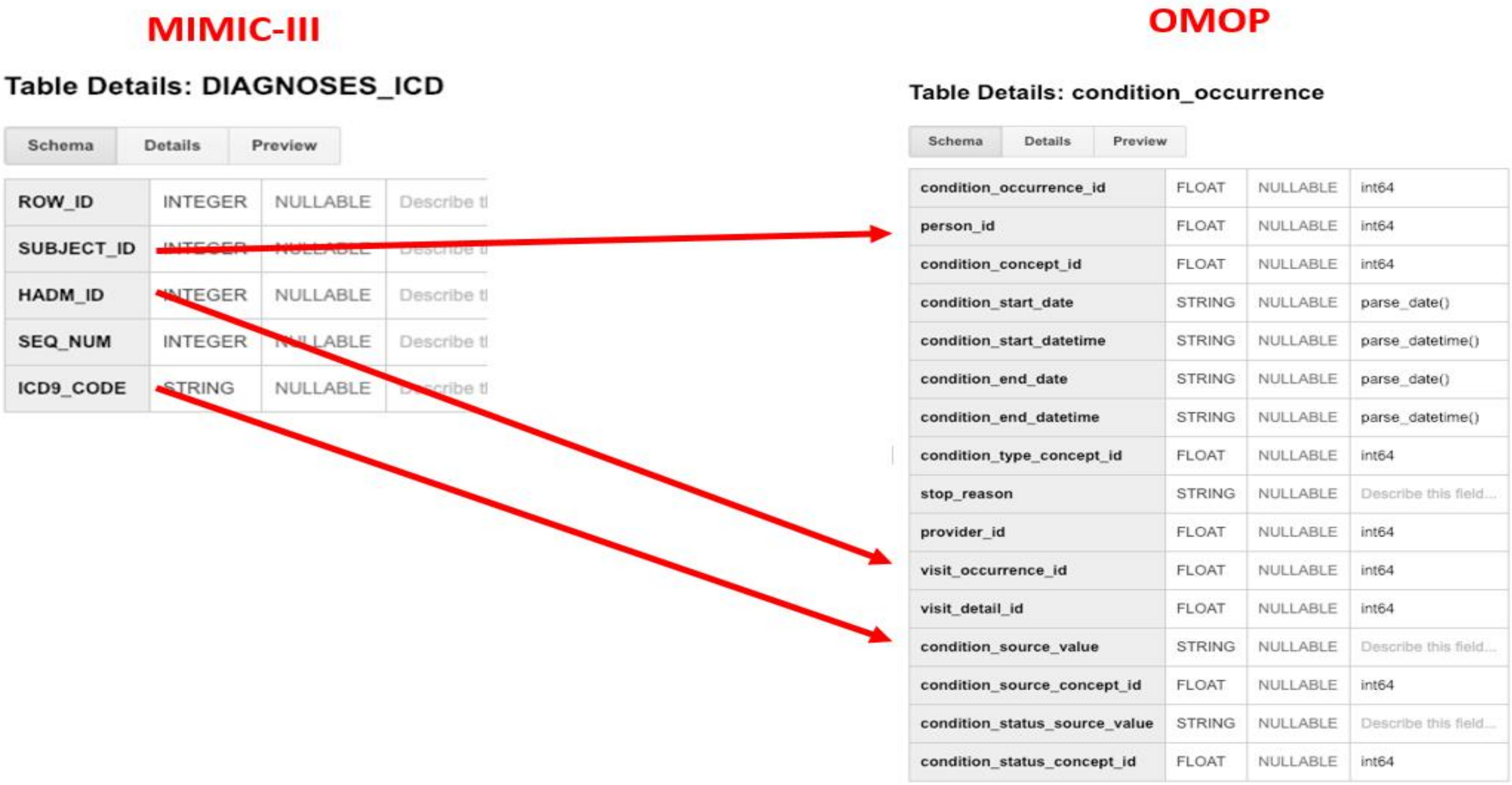


Table Details: TRANSFERS

ROW_ID	INTEGER	NULLABLE
SUBJECT_ID	INTEGER	NULLABLE
HADM_ID	INTEGER	NULLABLE
ICUSTAY_ID	INTEGER	NULLABLE
DBSOURCE	STRING	NULLABLE
EVENTTYPE	STRING	NULLABLE
PREV_CAREUNIT	STRING	NULLABLE
CURR_CAREUNIT	STRING	NULLABLE
PREV_WARDID	INTEGER	NULLABLE
CURR_WARDID	INTEGER	NULLABLE
INTIME	DATETIME	NULLABLE
OUTTIME	DATETIME	NULLABLE
LOS	FLOAT	NULLABLE

Use these screen captures (and previous slide) to select one or more MIMIC tables that contain data for OMOP CONDITION_OCCURRENCE table

Step 1: Understand source/target data models



MIMIC-III

Table Details: DIAGNOSES_ICD

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

OMOP

Table Details: condition_occurrence

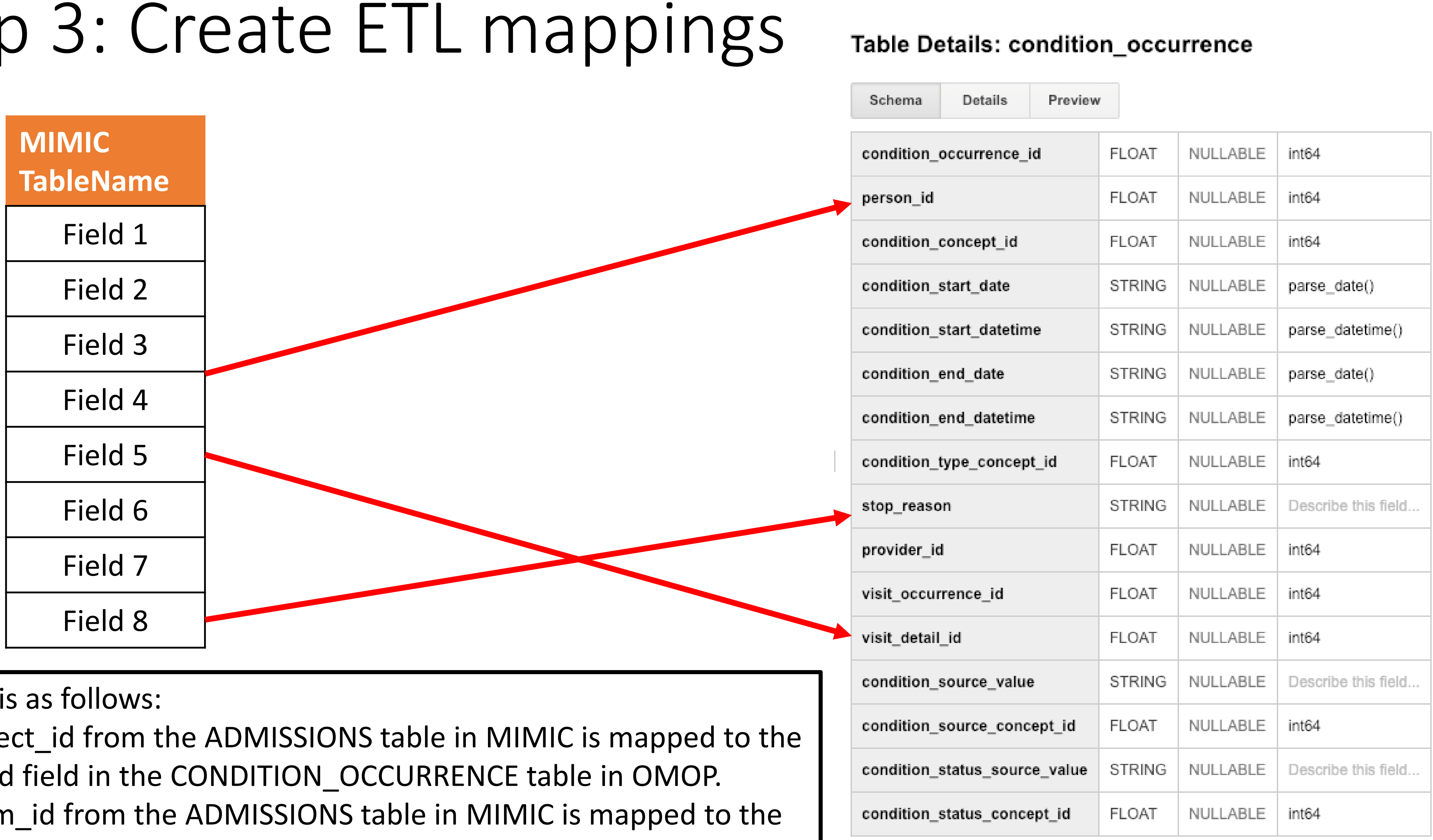
condition_occurrence_id	FLOAT	NULLABLE	int64
person_id	FLOAT	NULLABLE	int64
condition_concept_id	FLOAT	NULLABLE	int64
condition_start_date	STRING	NULLABLE	parse_date()
condition_start_datetime	STRING	NULLABLE	parse_datetime()
condition_end_date	STRING	NULLABLE	parse_date()
condition_end_datetime	STRING	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

- DIAGNOSES_ICD
  - SUBJECT_ID is a heavily right-skewed distribution, with just SUBJECT_ID 41976 accounting for about 15% of the available data in the report. A similar amount of records can be obtained with the next 6 most common patients, on which there exist 2–3% of records each. The rest of the patients are more uniformly represented in the table.

Step 3: Create ETL mappings



MIMIC TableName
Field 1
Field 2
Field 3
Field 4
Field 5
Field 6
Field 7
Field 8

Table Details: condition_occurrence

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

The mapping is as follows:

The subject_id from the ADMISSIONS table in MIMIC is mapped to the person_id field in the CONDITION_OCCURRENCE table in OMOP.

The hadm_id from the ADMISSIONS table in MIMIC is mapped to the visit_occurrence_id field in the CONDITION_OCCURRENCE table in OMOP

Step 4: Write transformation code

```
 WITH CONDITION_OCCURRENCE1 as (select
adm.subject_id as person_id, adm.hadm_id as
visit_occurrence_id
        from mimic3_demo.ADMISSIONS adm),

        CONDITION_OCCURRENCE as (select
oc1.person_id, oc1.visit_occurrence_id, diag.icd9_code
as condition_source_value
        from CONDITION_OCCURRENCE1 as oc1
        join mimic3_demo.DIAGNOSES_ICD as diag
        on oc1.visit_occurrence_id = diag.hadm_id)

select * from CONDITION_OCCURRENCE as oc
order by oc.person_id, oc.visit_occurrence_id;
```

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

Transformation code shown here is from the Course 2 videos showing transformation of MIMIC PATIENTS to OMOP PERSON

Step 5: Execute transformation code

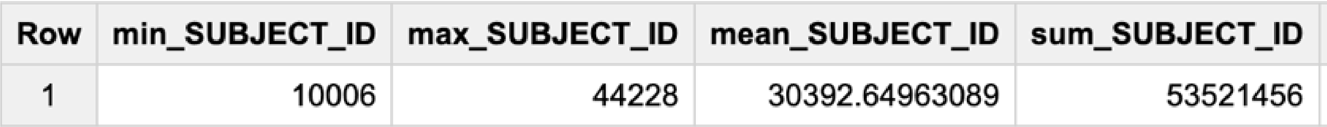
Execute the ETL code from Step 4 but do not submit the output table.
Use the output table for Step 6.

There is no submission for this Step.

Step 6: Perform data quality assessment

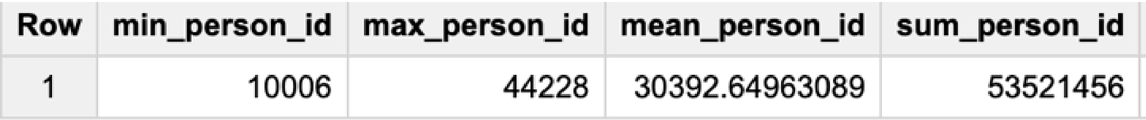
Define, implement, execute one or more data quality measures. Submit final DQ measure and an explanation why you created your measure(s).

MIMIC-III, SUBJECT_ID



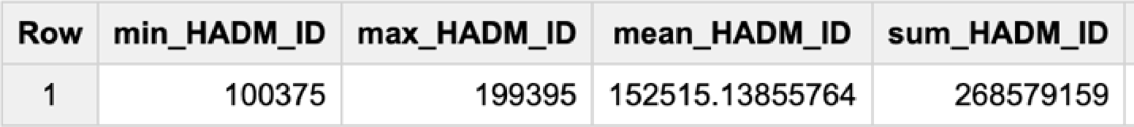
Row	min_SUBJECT_ID	max_SUBJECT_ID	mean_SUBJECT_ID	sum_SUBJECT_ID
1	10006	44228	30392.64963089	53521456

OMOP, person_id



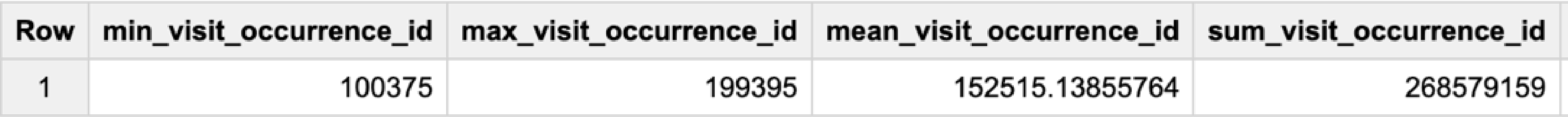
Row	min_person_id	max_person_id	mean_person_id	sum_person_id
1	10006	44228	30392.64963089	53521456

MIMIC-III, HADM_ID



Row	min_HADM_ID	max_HADM_ID	mean_HADM_ID	sum_HADM_ID
1	100375	199395	152515.13855764	268579159

OMOP, visit_occurrence_id



Row	min_visit_occurrence_id	max_visit_occurrence_id	mean_visit_occurrence_id	sum_visit_occurrence_id
1	100375	199395	152515.13855764	268579159

Step 7: Package documentation

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

There is no submission for this Step.