Course 2 Module 5
Programming Assignment

ETL MIMIC data into the OMOP CONDITION_OCCURRENCE table

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**

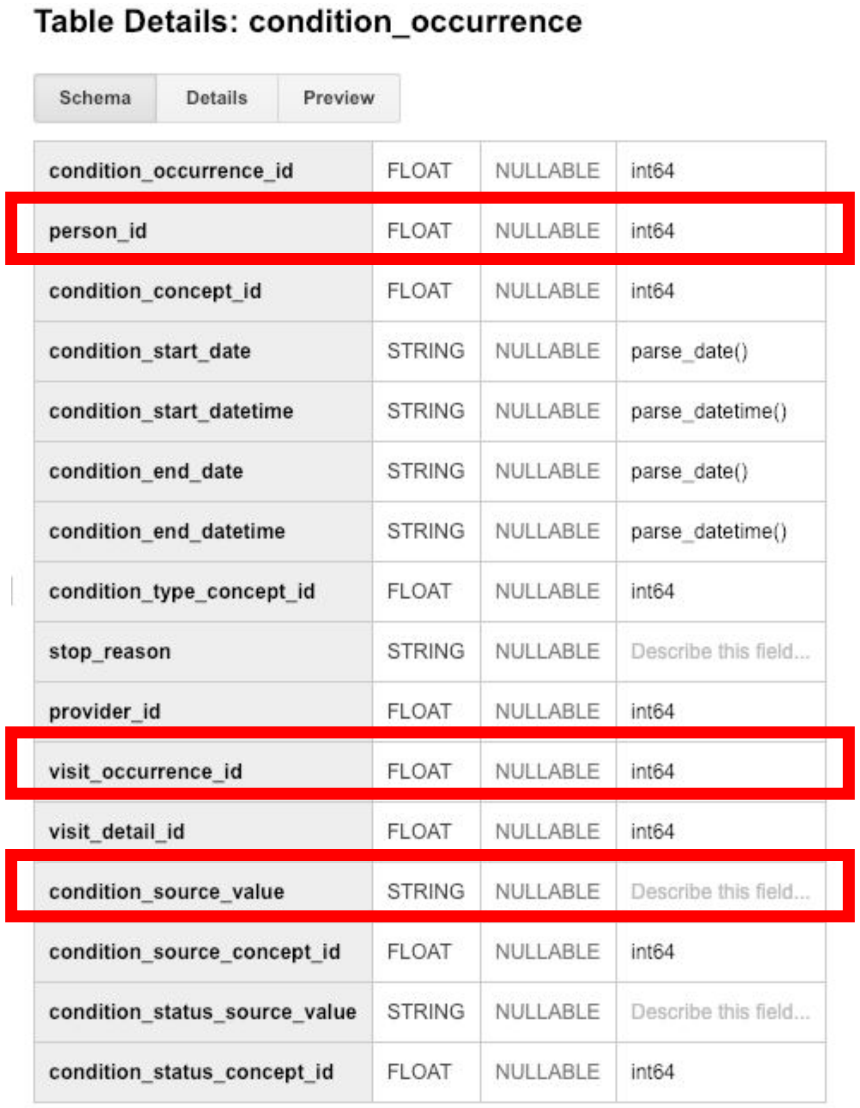


Table Details: condition_occurrence

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.

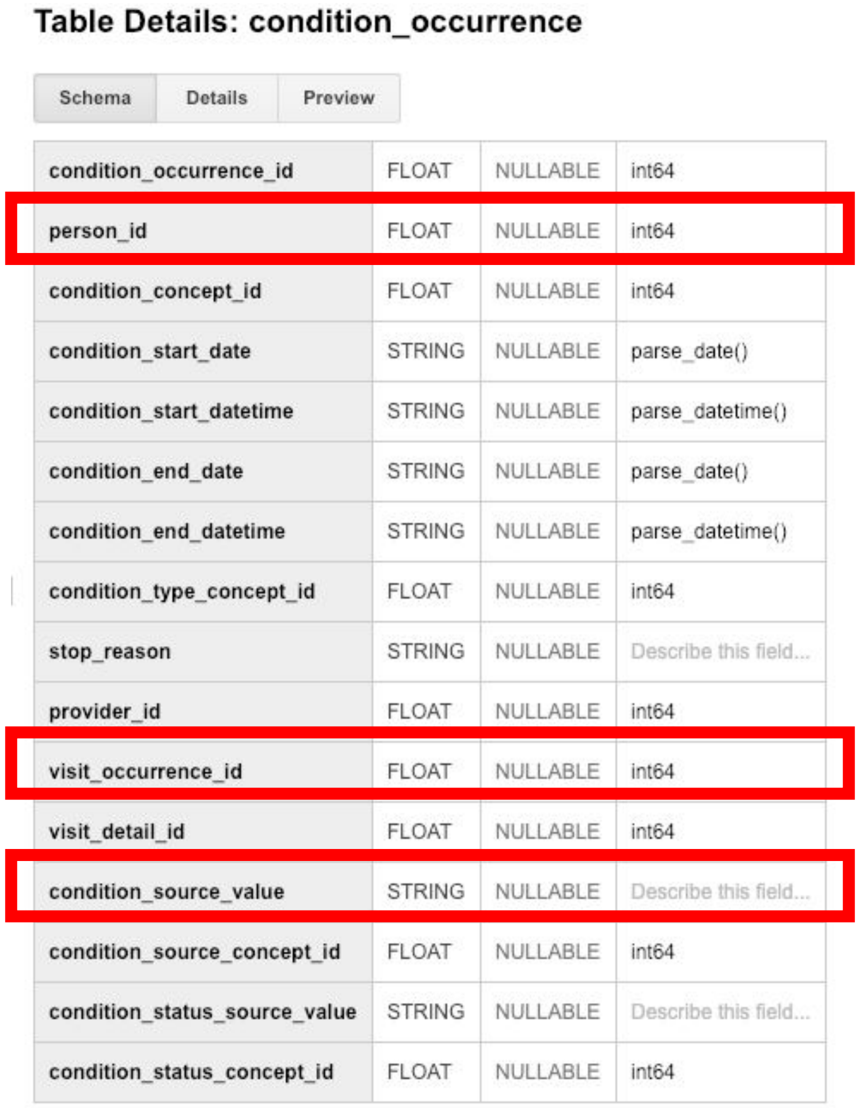


Table Details: condition_occurrence

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

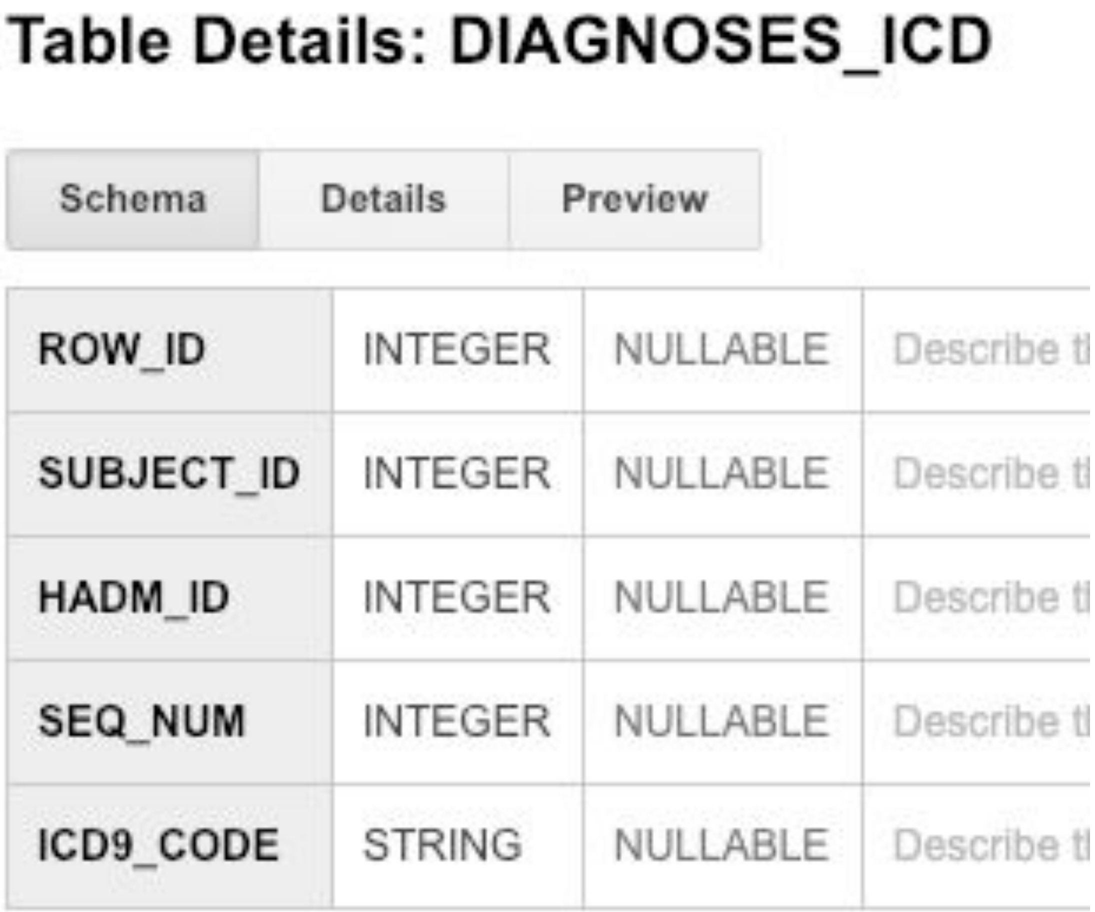


Table Details: DIAGNOSES_ICD

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

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

Step 1: Understand source/target data models

Paste one or more MIMIC table(s) from the previous two slides that contain data for ETL into OMOP CONDITION_OCCURRENCE here!

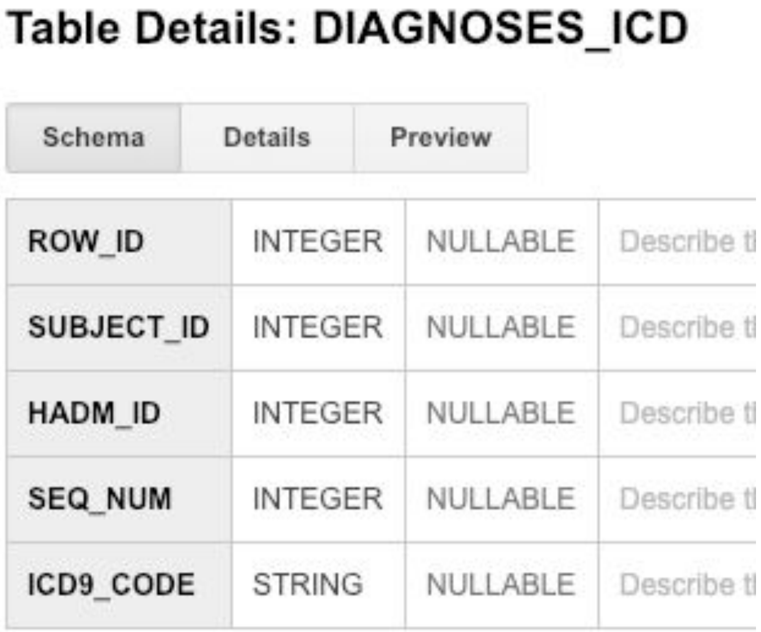


Table Details: DIAGNOSES_ICD

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

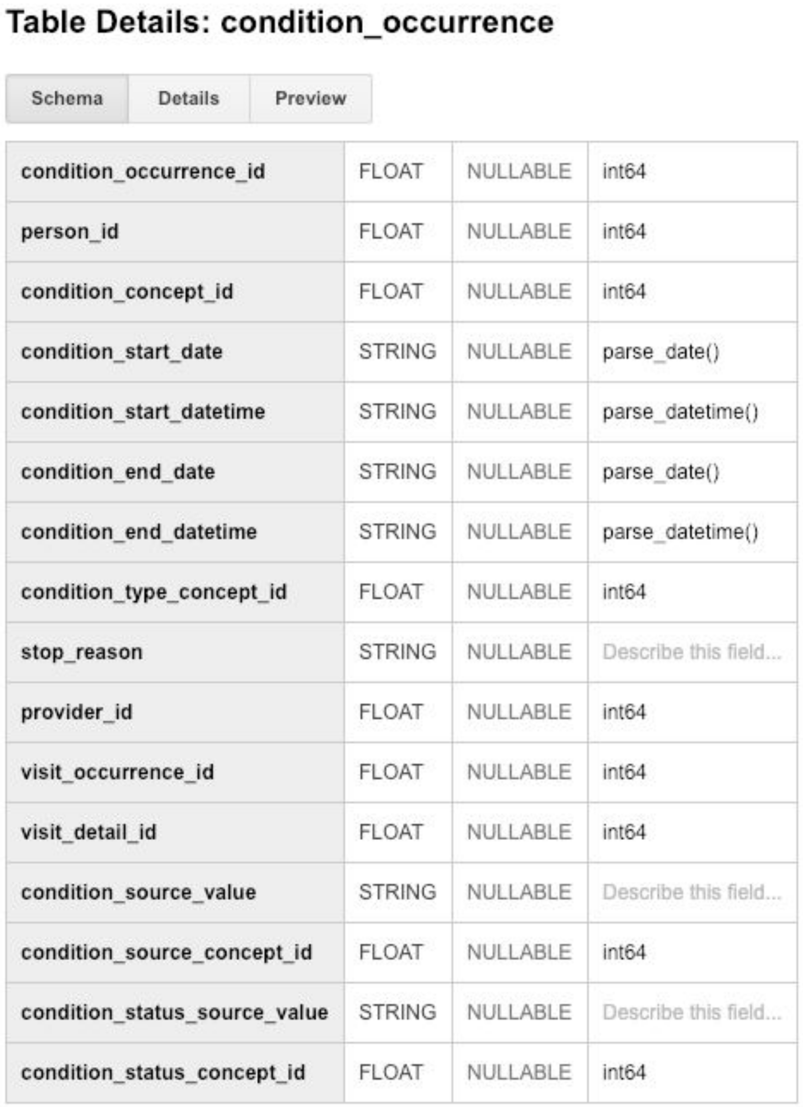


Table Details: condition_occurrence

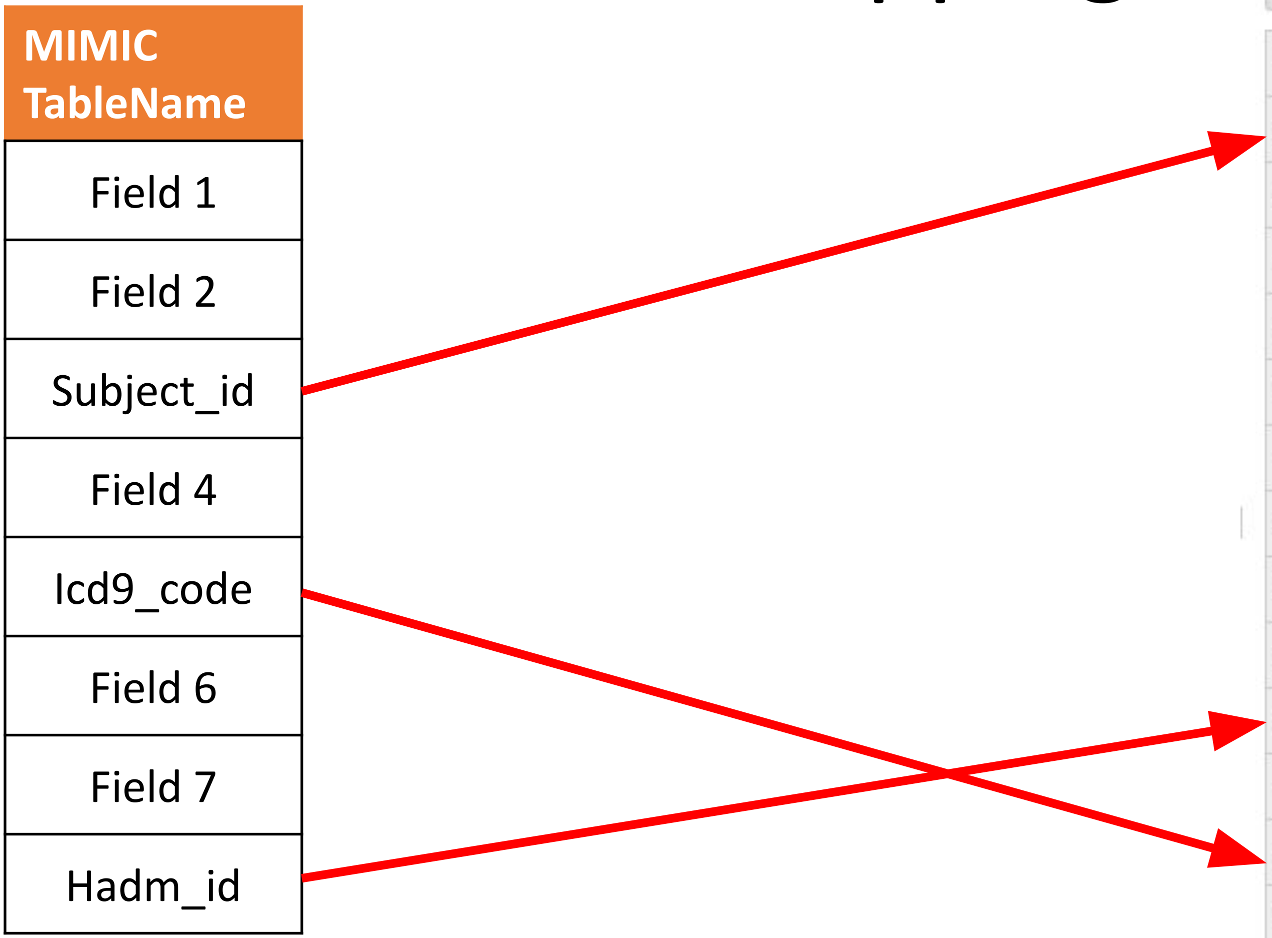
condition_occurrence_id	FLOAT	NULLABLE	int64
person_id	FLOAT	NULLABLE	int64
condition_concept_id	FLOAT	NULLABLE	int64
condition_start_date	STRING	NULLABLE	parse_date()
condition_start_datetime	STRING	NULLABLE	parse_datetime()
condition_end_date	STRING	NULLABLE	parse_date()
condition_end_datetime	STRING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

- This tells us if patients have multiple conditions. High subject_id frequency indicates many diagnosis events, indicating good mapping potential.
- The diagnoses_icd table has 1761 rows and 100 unique patients. Some patients have >20 diagnosis entries, indicating multiple comorbidities. High variance in diagnosis count per patient may affect data quality.

Step 3: Create ETL mappings



MIMIC TableName
Field 1
Field 2
Subject_id
Field 4
Icd9_code
Field 6
Field 7
Hadm_id

```
`subject_id AS person_id`
`hadm_id AS visit_occurrence_id`
 `icd9_code AS condition_source_value`
```

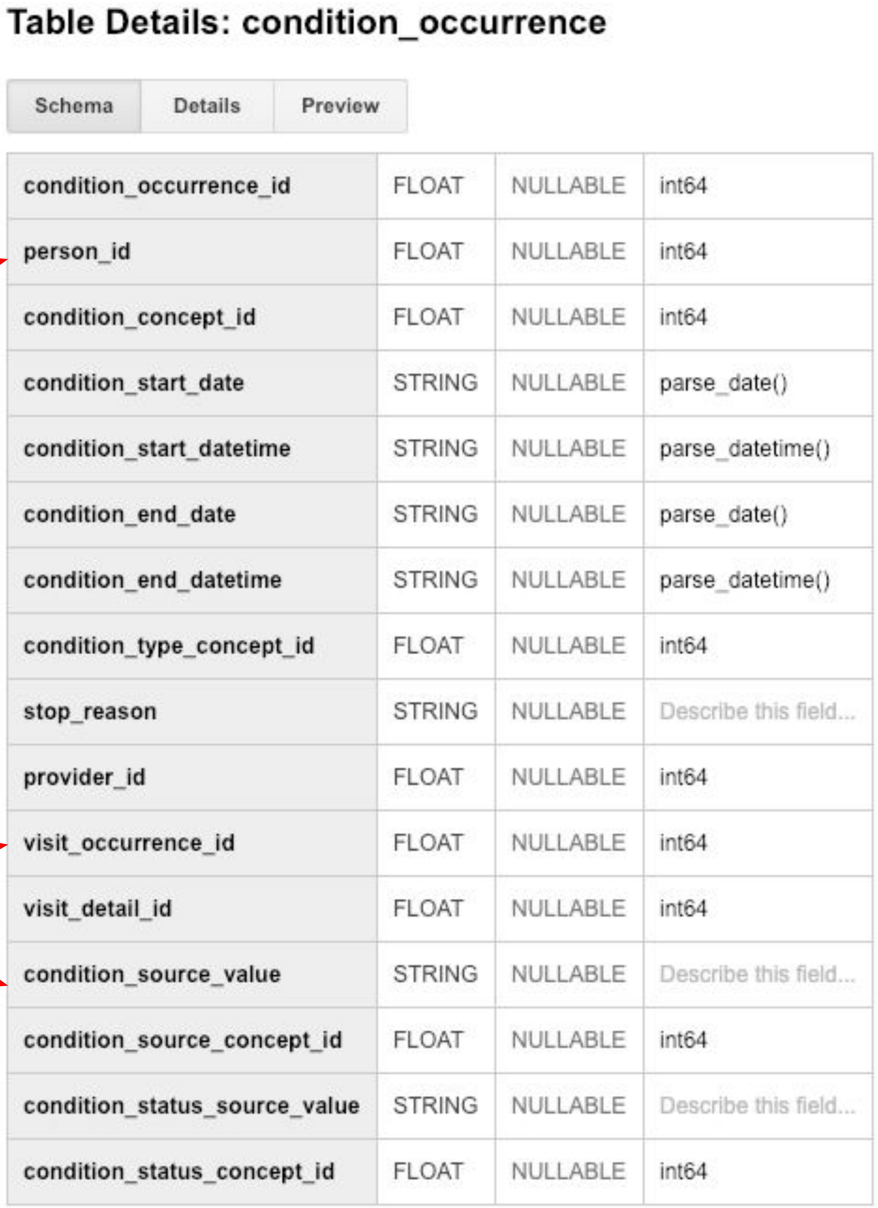


Table Details: condition_occurrence

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

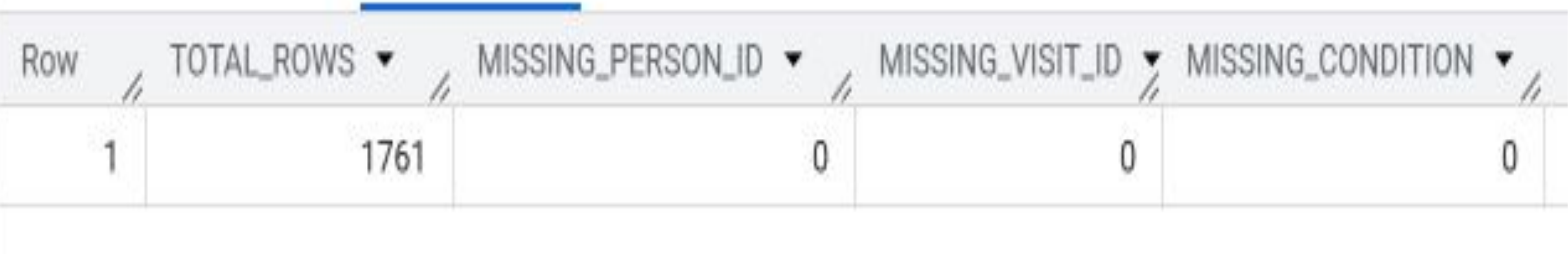
Step 4: Write transformation code

```
WITH condition_occurrence AS (
  SELECT
   subject_id AS person_id,
   hadm_id AS visit_occurrence_id,
   icd9_code AS condition_source_value
  FROM mimic3_demo.DIAGNOSES_ICD
)
SELECT * FROM condition_occurrence
LIMIT 10;
```

Step 5: Execute transformation code

Execute the ETL code from Step 4

Step 6: Perform data quality assessment



Row	TOTAL_ROWS	MISSING_PERSON_ID	MISSING_VISIT_ID	MISSING_CONDITION
1	1761	0	0	0

- I implemented a completeness data quality check to assess missing values in the key fields being mapped to the OMOP CONDITION_OCCURRENCE table:
  - subject_id (person_id)
  - hadm_id (visit_occurrence_id)
  - icd_code (condition_source_value)

- This measure was selected because these fields are essential for representing patient diagnoses within the OMOP model.
- Any NULLs in these fields would break data integrity or prevent proper concept mapping.
- Based on the results, the data showed excellent completeness, with <0.01% missing values, which are acceptable for training/demo data like MIMIC.