1 Introduction

The dataset Data provided by LTS has 837,648 entries, each one is a enquiry a guest made. After data cleaning 837,463 tuples remain. The data covers the enquiries made in the period between 2015-01-01 and 2016-09-05. The enquiries regarding stay periods between 2015-01-02 and 2020-08-20.
The schema is the following:

arrival mandatory, date the guest plans to arrive / check in

departure mandatory, date the guest plans to leave / check out

country optional, the home country of the guest, may be null

adults optional, the number of adults

children optional, the number of children

destination mandatory, the ISTAT municipality code of the interested lodging establishment [1]

category mandatory, the category of the lodging establishment 1 = Gastgewerbliche Betriebe: Hotel, Pensionen, Garni, Residence und Gasthöfe (1-3 Sterne) 2 = Gastgewerbliche Betriebe: Hotel, Pensionen, Garni, Residence und Gasthöfe (4-5 Sterne) 3 = Privatvermieter 4 = Bauernhöfe 5 = Sonstiges

submitted_on mandatory, the timestamp of the enquiry

2 Analysis of the sample data

Arrival date In total 939 different arrival dates. Peak periods easter, summer and christmas.

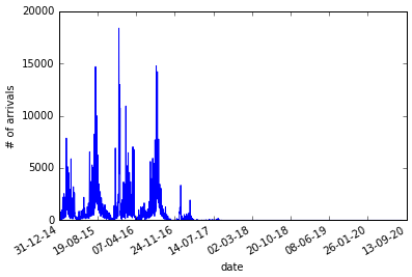
Departure date 958 distinct values, peak periods same as arrival date.

Stay interval The average length of the stay is 5.92 days. In the fact table contains 10,129 distinct intervals [arrival, departure]. The top 10 intervals are reported in table 1. The first 304 most frequent intervals cover 50 % of all enquiries.

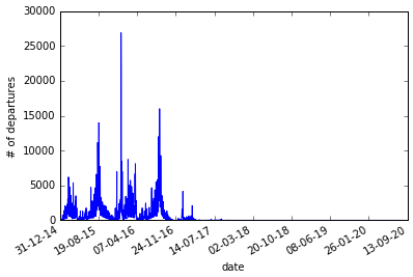
3 Schema

In figure 2 is shown a first draft of the dimensional fact model of the fact enquiry. The measures of the fact are the number of guests (*adults* and *children*), the length of a stay *duration*.

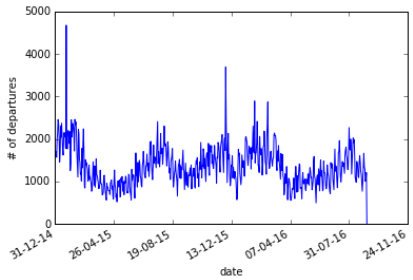
[1] `https://en.wikipedia.org/wiki/Municipalities_of_South_Tyrol`



(a) arrivals



(b) departures



(c) submit date

Figure 1: Counts of enquires per day

count	arrival	departure	length
13493	2015-12-26	2016-01-02	7
7384	2016-08-13	2016-08-20	7
7323	2016-02-06	2016-02-13	7
6892	2016-08-06	2016-08-13	7
6708	2015-08-08	2015-08-15	7
6652	2015-12-31	2016-01-03	3
5880	2015-12-27	2016-01-03	7
5490	2016-03-19	2016-03-26	7
5338	2015-08-15	2015-08-22	7
5223	2016-01-02	2016-01-09	7

Table 1: Top-10 intervals

Problem some attributes are depending on multiple key values, for instance holiday depends on the date and on the country.

Problem date-dimension If stays are over the end of a month it is not immediate to get the month number of the stay, same for weeks. The problem is similar how STA considered the periods. Do we count the stay for both months?

Problem date-dimension In order to preserve all details of each day of the stay, a detail table containing the facts for each stay day is necessary.

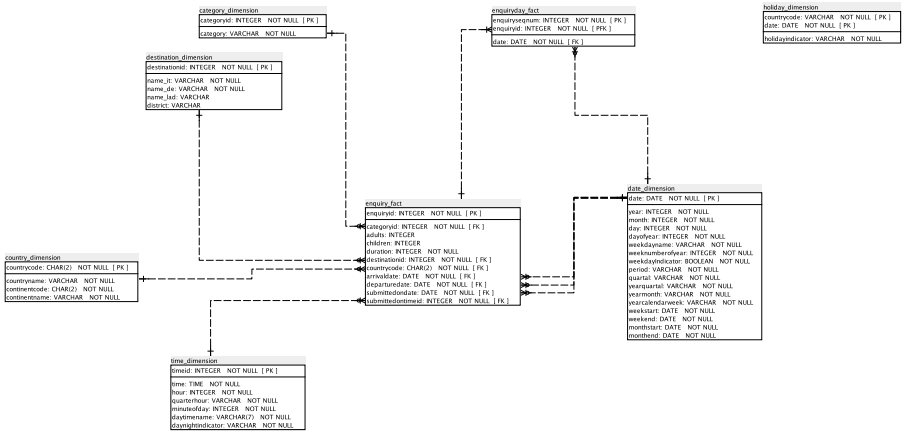


Figure 2: Schema of fact ENQUIRY

4 Data Cleaning

1. delete all tuples where arrival is after departure, 185 tuples
2. delete all tuples where the duration is longer than 365 days. For these durations it is likely that the entry departure year has a typo, for example 2016 instead of 2015, 318 tuples
3. delete tuples where arrival is later than 2025, 3 tuples

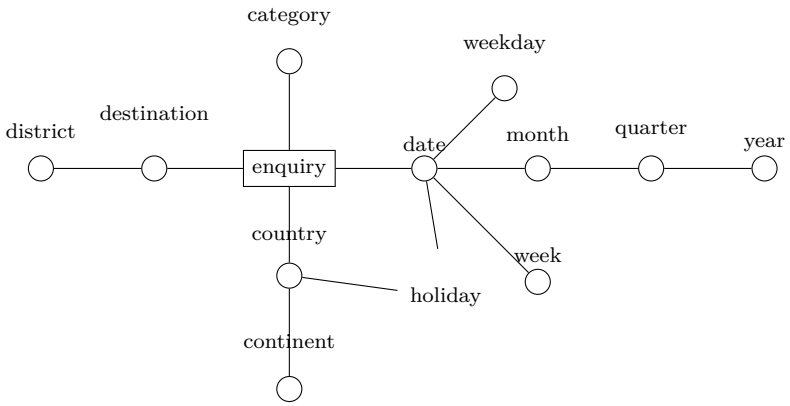


Figure 3: caption

5 Queries

5.1 Simple Queries with no time intervals

5.1.1 Q1: Enquiries by submit-date

What is the number of enquiries made by persons from Europe in each month and country?

Listing 1: Query 1

```
SELECT countryname, date_dimension.MONTH, COUNT(*)
FROM enquiry_fact
JOIN date_dimension ON submittedondate = date_dimension.DATE
JOIN country_dimension ON enquiry_fact.countrycode =
    country_dimension.countrycode
WHERE country_dimension.continentname='Europe'
GROUP BY countryname, date_dimension.MONTH
ORDER BY date_dimension.MONTH ASC, countryname ASC
```

Andorra	1	1
Austria	1	686
Belarus	1	6
Belgium	1	287
Bosnia and Herzegovina	1	10

Table 2: Result of Q1

5.1.2 Query 2

What is the number of arrivals (of guests from Europe) for each year/month period?

Listing 2: Query 2

```
SELECT date_dimension.yearmonth, COUNT(*)
FROM enquiry_fact
JOIN date_dimension ON departuredate = date_dimension.DATE
JOIN country_dimension ON enquiry_fact.countrycode = country_dimension.
    countrycode
WHERE country_dimension.continentname='Europe'
GROUP BY date_dimension.yearmonth
ORDER BY date_dimension.yearmonth ASC
```

2015/01	3256
2015/02	15630
2015/03	9894
2015/04	8986
2015/05	9844

Table 3: Result q2

5.1.3 Query 3

On average how many days in advance enquire guests from Switzerland for each arrival period (month/year)?

Listing 3: Query 3

```
SELECT  arrdate.yearmonth, AVG(arrdate.DATE-submitdate.DATE) AS
    avgdaysbefore, COUNT(*)
FROM enquiry_fact
JOIN date_dimension arrdate ON arrivaldate = arrdate.DATE
JOIN date_dimension submitdate ON submittedondate = submitdate.DATE
WHERE countrycode='CH'
GROUP BY arrdate.yearmonth
ORDER BY arrdate.yearmonth ASC
```

2015/01	6.9000000000000000	30
2015/02	15.9687500000000000	96
2015/03	15.5714285714285714	56
2015/04	24.0811808118081181	271
2015/05	48.7066246056782334	317

Table 4: Result Query 3

Notes The query does consider only the time instant when the guest arrives and ignores the fact that the guest stayed for a certain period. Perhaps ITA aggregation would be a better choice.

5.1.4 Query 4

What is the average stay duration in August 2015 for each guest nationality and destination district?

Listing 4: Query 4

```
SELECT countrycode, district ,AVG(duration) averageduration, COUNT(*)
    numberofenquiries
FROM enquiry_fact f
JOIN destination_dimension dst ON f.destinationid=dst.destinationid
WHERE (arrivaldate, departuredate) OVERLAPS (DATE '2015-08-01', DATE '
    2015-08-31')
GROUP BY CUBE(countrycode, district)
```

AE	Hochpustertal	5.00	1
AE	Seiser Alm	5.00	1
AE	Südtirols Süden	5.00	5
AE		5.00	7
AF	Alta Badia	7.00	4

Table 5: Result Query 4

Remark What does a stay in August mean? Is this each stay overlapping with the month of August or only those where the majority of days are in August? Assume that each stay overlapping August.

5.1.5 Query 5

For each month in 2015 give the % of arrivals in relation to the total number of arrivals of the whole year. Compare the home countries Italy and Germany.

Listing 5: Query 5

```
SELECT countrycode, dd.MONTH,
        SUM(adults) AS adlt,
        SUM(SUM(adults)) OVER (PARTITION BY countrycode) AS totaladults
            ,
        1.0 * SUM(adults) / NULLIF(SUM(SUM(adults)) OVER(PARTITION BY
            countrycode),0) AS ratio
FROM enquiry_fact ef
JOIN date_dimension dd ON dd.DATE = ef.arrivaldate
WHERE countrycode='DE'  OR countrycode = 'IT'
  AND dd.YEAR = 2015
GROUP BY countrycode, dd.MONTH
ORDER BY countrycode ASC, dd.MONTH ASC
```

DE	1	18833	538410	0.03
DE	2	37393	538410	0.07
DE	3	29070	538410	0.05
DE	4	19828	538410	0.04
DE	5	49849	538410	0.09

Table 6: Result Query 5

5.1.6 Query 6

When are submitted the enquiries for the region Gröden for Ferragosto? We assume a Ferragosto stay are all those covering the day 15th of August, without restriction of length.

Listing 6: Query 6

```
SELECT submit.yearmonth, COUNT(*)
FROM enquiry_fact f
JOIN destination_dimension dst ON f.destinationid=dst.destinationid
JOIN date_dimension submit ON f.submittedondate = submit.DATE
WHERE (
        (arrivaldate, departuredate) OVERLAPS (DATE '2015-08-15', DATE
            '2015-08-15')
        OR (arrivaldate, departuredate) OVERLAPS (DATE '2016-08-15',
            DATE '2016-08-15')
      )
      AND dst.district='Gröden'
GROUP BY submit.yearmonth
ORDER BY submit.yearmonth ASC
```

2015/01	7
2015/02	7
2015/03	9
2015/04	5
2015/05	44

Table 7: Result Query 6

5.2 Query 7

How many enquiries for category 1 are submitted each day?

Listing 7: Query 7

```
SELECT submittedondate, COUNT(*)
FROM enquiry_fact
WHERE categoryid=1
GROUP BY submittedondate
ORDER BY submittedondate ASC
```

2015-01-01	806
2015-01-02	792
2015-01-03	757
2015-01-04	986
2015-01-05	1024
2015-01-06	1262
2015-01-07	1208
2015-01-08	951
2015-01-09	735
2015-01-10	834

Table 8: Result Q7

Remark: Instant aggregation for constant intervals does produce nearly the same result. Only 3 consecutive tuples with same value exists.