DMIF, University of Udine

Data Management for Big Data

The Gap Srlu Case

Andrea Brunello
andrea.brunello@uniud.it

April 2022

Outline

Introduction: The Contact Center Domain

Multi-channel contact centers are an important component of today's business world.

They serve as a primary customer-facing channel for firms in many different industries, and employ millions of agents across the globe.

During their operation, they generate vast amounts of heterogeneous data, ranging from structured automatically registered logs to semi-structured hand-written notes and unstructured raw voice recordings.

Inbound, Outbound and Backoffice Ops.

Inbound contact centers handle incoming traffic, e.g., they answer to calls received from the customers, as in the case of help-desks.

Outbound contact centers handle outgoing traffic, which is initiated from the center. For example, this is the case of calls associated with surveys or telemarketing initiatives, that typically follow a predefined script.

Backoffice operations may also be carried out, as in the case of data preparation and data analysis tasks.

All operations are carried out within the context of a *service* (e.g., an airline toll-free number), which can be composed of many different *activities* (e.g., ticket booking, or car rental).

Gap Srlu Company

The Company

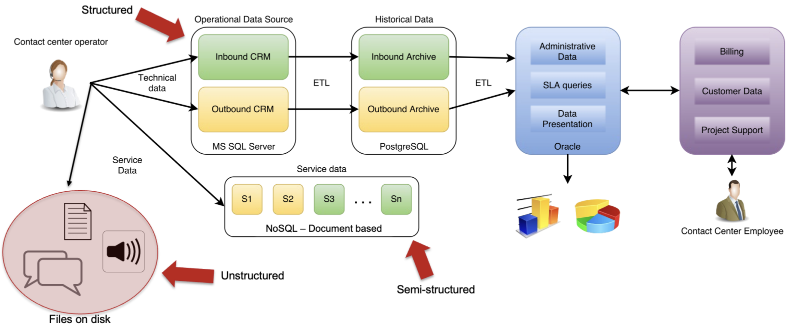
Gap Srlu is a multi-channel and multi-service Business Process Outsourcer, specialized in contact center activities.

It is active since the early 2000s and, over time, it has experienced a continuous expansion concerning both its business model and its information system infrastructure.

Nowadays, other than the traditional contact center tasks, it is capable of offering advanced services such as third-party data management and analysis, based on several machine learning technologies.

More info at: *https://www.gapitalia.com/en/*

The Initial Situation



What are the Issues Here?

Several problems:

- heterogeneous systems require ad-hoc solutions for reading and writing data
- different databases adopt different conventions for storing the data
- possibly (and probably) replicated and inconsistent information
- difficult to perform queries and analyses involving more than one data repository
- some of the data are not even considered for analytics purposes
- the whole architecture is complex, and hard to maintain and update

Development of the Data Warehouse

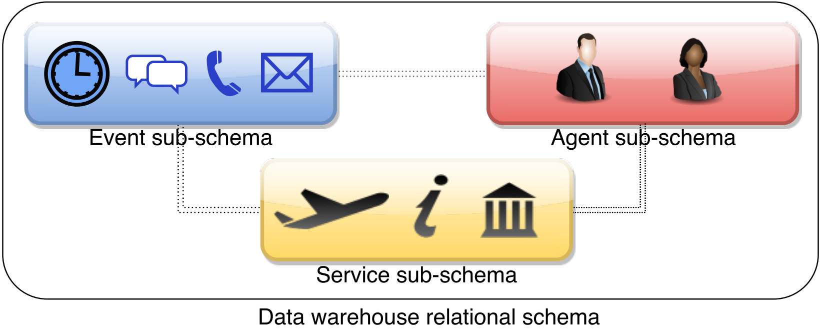
Why a Data Warehouse

All kind of monitoring and analysis tasks start from the data.

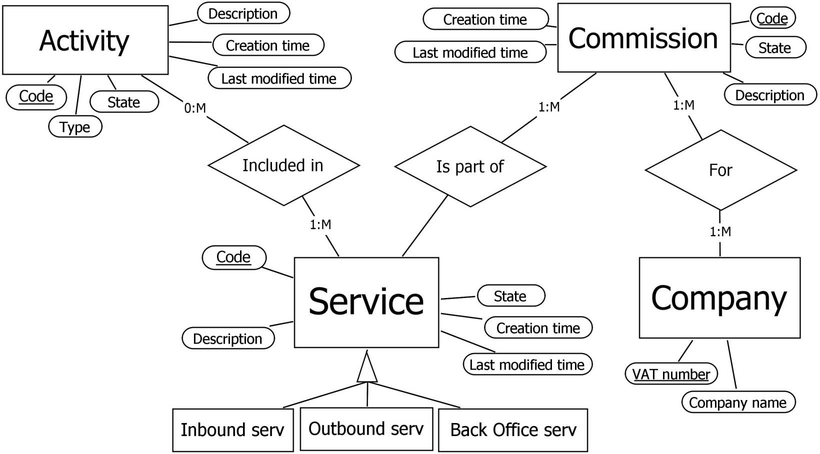
Thus, there is the necessity of having a clear and uniform view over all the company information.

Moreover, a unique, central data repository simplifies the overall infrastructure.

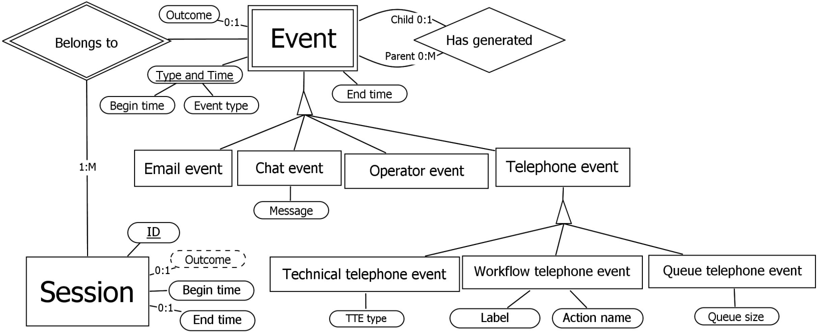
Data Warehouse Overall Design



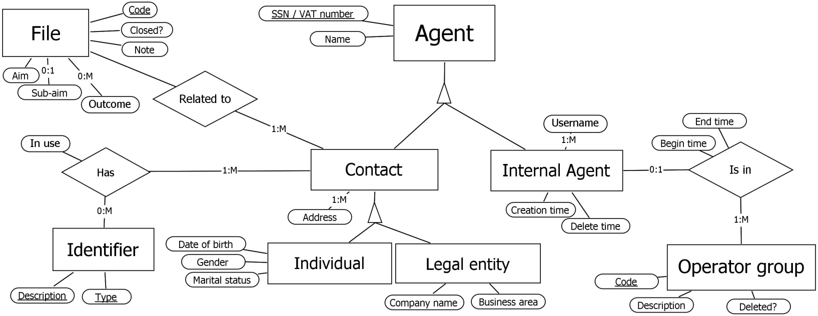
Service Sub-schema



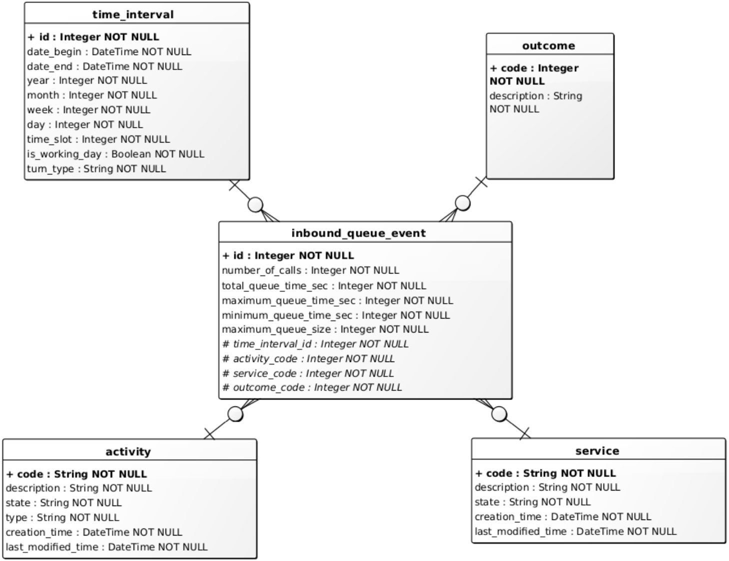
Event Sub-schema



Agent Sub-schema



The Analysis Layer / Data Marts



Analysis Tasks

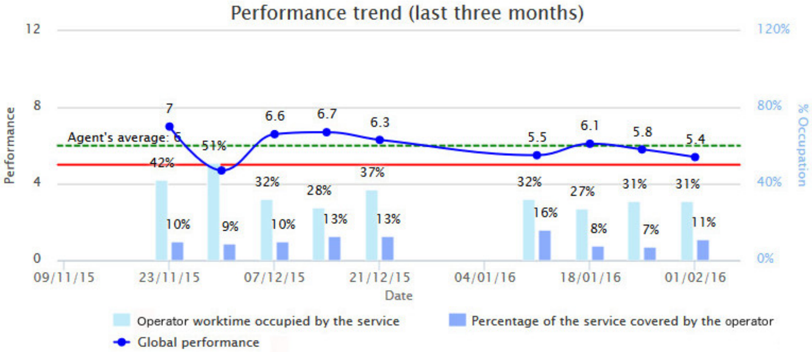
Operator Performance Assessment

Tracking the performance of agents is a primary issue in contact centers, as it allows, for example:

- the best match to be taken between a service and an agent
- the recognition of unsatisfactory agent behaviours, due for example to a lack of proper training
- the prediction of future trends, based on the history of observations

A function has been designed, which is capable of assigning a score to each operator-service couple.

Inbound	Average conversation time
	Average postcall time
	Generic call notes compiled per session
	Percentage of correctly filled script fields
	Purpose of the call
	Outcome of the call
Outbound	Average conversation time
	Average postcall time
	Amount of surveys over calls
	Number of answered calls per hour
General	Number of different kinds of services managed by an operator
	Degree of interleaving between services
	Respect of work schedule
	Turn flexibility


Performance trend (last three months)
Performance
% Occupation
Agent's average: 6
7
6.6
6.7
6.3
5.5
6.1
5.8
5.4
42%
51%
32%
28%
37%
32%
27%
31%
31%
10%
9%
10%
13%
13%
16%
8%
7%
11%
12
8
4
0
120%
80%
40%
0%
09/11/15
23/11/15
07/12/15
21/12/15
04/01/16
18/01/16
01/02/16
Date
Operator worktime occupied by the service
Percentage of the service covered by the operator
Global performance

Analysis of Written Notes

As a part of the agent performance evaluation framework, Gap automatically assesses the characteristics of written notes taken by the agents during phone calls:

- how often / in which way does an agent record notes regarding an inbound call?
- compare single agent behaviour with service average values

How to evaluate written notes?

- extract summarizing features from the text
- identify groups of similar notes
- devise a methodology to assign a generic new note to one of the previously identified groups

For each note, we calculate:

- numbers of words and characters
- *Gulpease* readability index value
- fractions of articles and conjunctions over words
- fractions of verbs and adverbs over words
- fraction of adjectives over words
- fraction of prepositions over words
- fraction of quantifiers over words
- fraction of (pro)nouns over words
- fraction of numeric codes over words
- fraction of proper nouns over words
- fraction of words/abbreviations found in Italian dictionary
- fraction of words found in **service-specific** domain (Wiki)
- fraction of unrecognized words

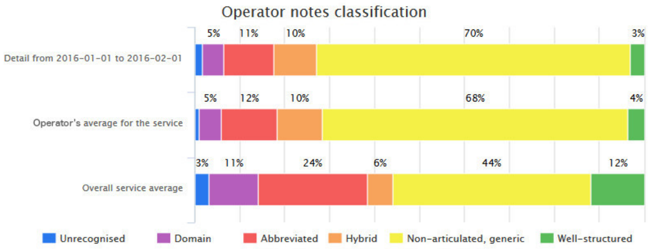
- Random sampling of 1000 notes
- application of a clustering algorithm to the selected notes (*E-M* algorithm)
- 6 clusters emerged:
  - articulated notes
  - non-articulated notes
  - abbreviated notes
  - domain-specific notes
  - nonsense notes
  - hybrid notes

- Attach a new feature to each of the clustered notes: *cluster label*
- apply a decision tree learning algorithm (*J48*), with the goal of predicting the label (94.7% accuracy)
- the tree can then be used to classify new notes

```
riconosciuti_abbr_su_parole <= 0.142857
|   riconosciuti_dominio_su_parole <= 0.133333
|   |   preposizioni_su_parole <= 0
|   |   |   non_riconosciuti_su_parole <= 0.157895
|   |   |   |   congiunzioni_su_parole <= 0.025
|   |   |   |   |   articoli_su_parole <= 0.071429: non_articulated_notes
|   |   |   |   |   articoli_su_parole > 0.071429: articulated_notes
|   |   |   |   congiunzioni_su_parole > 0.025: articulated_notes
|   |   |   non_riconosciuti_su_parole > 0.157895
|   |   |   |   non_riconosciuti_su_parole <= 0.333333
|   |   |   |   |   articoli_su_parole <= 0.083333: hybrid_notes
|   |   |   |   |   articoli_su_parole > 0.083333: articulated_notes
|   |   |   |   non_riconosciuti_su_parole > 0.333333: non_sense_notes
|   |   preposizioni_su_parole > 0
|   |   |   indice_gulp <= 129.833333: articulated_notes
|   |   |   indice_gulp > 129.833333
|   |   |   |   non_riconosciuti_su_parole <= 0.0625: non_articulated_notes
|   |   |   |   non_riconosciuti_su_parole > 0.0625: hybrid_notes
```

	valore text	gruppo_nota text
1	info voltura	hybrid
2	invio del f24 [illegible]	articulated
3	informazioni per appunt sub e comunica dati catastali	articulated
4	info posizione pagamenti mensa scolastica	hybrid
5	NON RISPONDE	non-articulated
6	Info	abbreviated
7	VIA [illegible] MQ 37 C'è SCRITTO 43 BOLLETTAZIONE SBAGLIATA. DEVE PASSARE AGLI SPORTELLI PER RETTIFICA DI METRATURA CON PIANTINA SCALA 1:100. RIFERISCO. C	articulated
8	SIGNORA CHIAMA PER SAPERE SE è STATA APPLICATA LA DETRAZIONE DI 25 euro per figlio sul calcol	articulated
9	la signora avea chiamato il 23/05 per una verifica posizione per la TARES: ha un locale comme	articulated
10	chiede quanto deve pagare per la tassa. Parlato con [illegible]: deve pagare 61 euro.	articulated
11	info boll	abbreviated
12	rimborso ud	non-articulated
13	tasi	domain-specific
14	INFO GENERICHE IMU, TASI	domain-specific
15	info su avv sosp	hybrid
16	chiede se può rateizzare l'importo da versare per la mensa. Riferito che deve fare richiesta	articulated
17	invio copia boll	hybrid
18	chiede il saldo mensa. Riferito che abbiamo problemi tecnici tecnici al server	articulated

Agent-service notes class distribution, with respect to the overall distribution for the service.



Anomalous Call Outcomes Detection

Outbound calls follow a pre-defined script, which allows one to predict, to a certain extent, their outcome based just on *dialling*, *conversation*, and *postcall* times.

This allows to detect contact center operators who systematically annotate wrong call outcomes, either by mistake or to simulate surveys which did not take place.

A decision tree model has been developed that, based on *dialling*, *conversation*, and *postcall* times of a phone conversation, derives its most likely outcome, with an accuracy above 93%.

The training set was composed of phone calls performed by trustworthy operators.

```
conversation_time <= 7
|   conversation_time <= 0
|   |   dialling_time <= 30
|   |   |   dialling_time <= 11: busy_or_nonexistent
|   |   |   dialling_time > 11
|   |   |   |   dialling_time <= 14: busy_or_nonexistent
|   |   |   |   dialling_time > 14: no_answer
|   |   dialling_time > 30: no_answer
|   conversation_time > 0
|   |   postcall_time <= 1
|   |   |   dialling_time <= 29: fax_or_answermachine
|   |   |   dialling_time > 29
|   |   |   |   conversation_time <= 1: no_answer
|   |   |   |   conversation_time > 1: fax_or_answermachine
|   |   postcall_time > 1
|   |   |   conversation_time <= 4: fax_or_answermachine
|   |   |   conversation_time > 4: spoken_no_survey
conversation_time > 7
|   conversation_time <= 76
|   |   conversation_time <= 11
|   |   |   postcall_time <= 1
|   |   |   |   conversation_time <= 9
|   |   |   |   |   dialling_time <= 22
|   |   |   |   |   |   conversation_time <= 8: fax_or_answermachine
|   |   |   |   |   |   conversation_time > 8: spoken_no_survey
|   |   |   |   |   dialling_time > 22: fax_or_answermachine
|   |   |   |   conversation_time > 9: spoken_no_survey
|   |   |   postcall_time > 1: spoken_no_survey
|   |   conversation_time > 11: spoken_no_survey
|   conversation_time > 76
|   |   conversation_time <= 87
|   |   |   postcall_time <= 0: spoken_no_survey
|   |   |   postcall_time > 0: survey_made
|   |   conversation_time > 87: survey_made
```

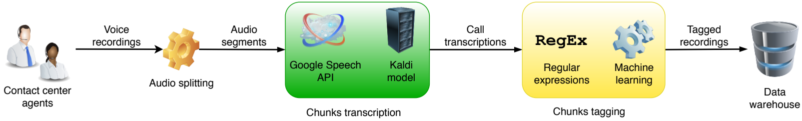
The ability to analyze conversational data plays a major role in contact centers, where the core part of the business still focuses on the management of oral interactions.

Several actors already provide speech analytics solutions, e.g., Google or Amazon. However, they come with a price.

Is it possible to develop an in-house effective speech analytics framework in a cost-effective manner, considering the Italian language?

The focus is on agent voice recordings generated within an outbound survey.

The content of the recordings is typically not too heterogeneous (due to the presence of a script).



An in-house speech-to-text model has been developed, based on the framework Kaldi (*https://kaldi-asr.org/*) and the following corpora.

Corpus name	**# utterances**		**Recording time**	
	training	**test**	**training**	**test**
CLIPS	1025	-	2h 30m	-
QALL-ME	1208	-	2h 20m	-
Proprietary (read)	3467	-	4h 28m	-
Proprietary (spontaneous)	201	339	30m	35m

A word error rate of 28.77% was achieved, compared to 18.70% which could be obtained relying on Google Cloud Speech API. This is enough to perform some analyses over the transcripts.

Several kinds of analysis tasks may be performed over the obtained textual data.

For instance, it is possible to determine whether the agent has pronounced all the parts required by the script (for instance: introduction, script question #1, privacy statement, . . .).

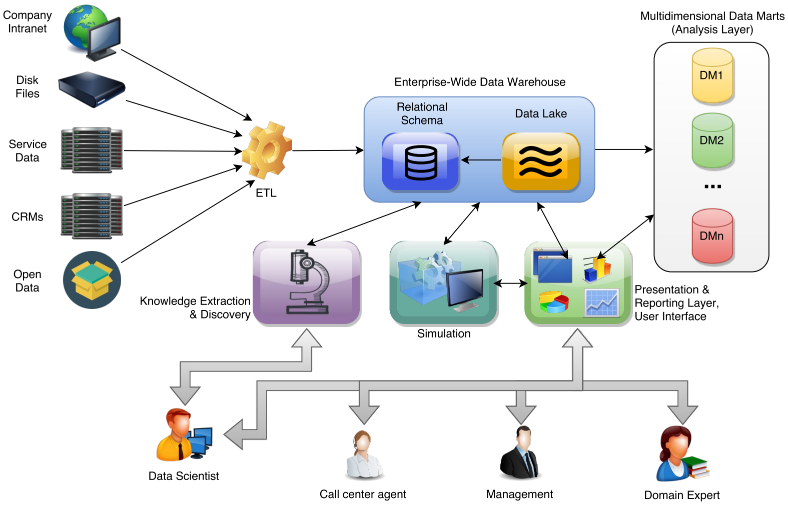
The overall idea is that of attaching tags to the transcribed phrases, in order to track the presence of different script parts.

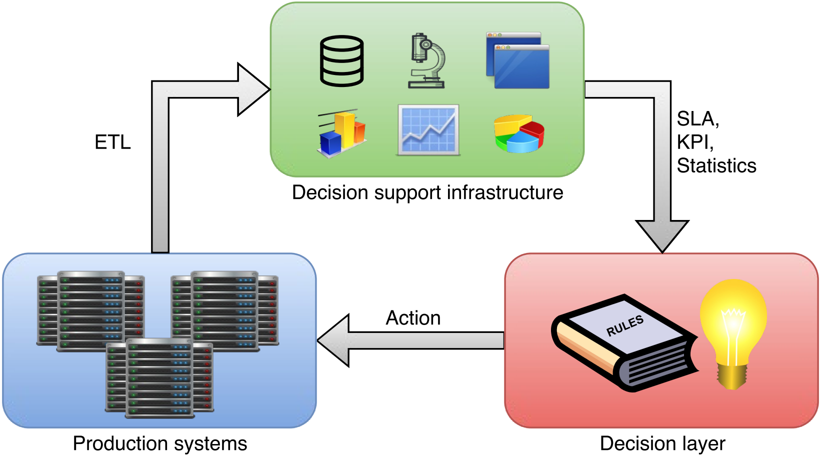
This can be done based on user-defined regular expressions, or using some more advanced machine learning approaches.

Performance obtained by several approaches, on the task of tag identification in the call transcripts.

	Accuracy		Precision		Recall		TNR	
Keyword	**K**	**G**	**K**	**G**	**K**	**G**	**K**	**G**
Regular expressions	0.966	0.942	0.912	0.928	0.763	0.575	0.990	0.992
Logistic, unigram	0.972	0.973	0.903	0.916	0.839	0.870	0.989	0.973
Logistic, bigram	0.961	0.966	0.917	0.923	0.691	0.789	0.992	0.980
Logistic, trigram	0.940	0.951	-	0.910	0.494	0.666	0.995	0.895
Hybrid	0.974	0.973	0.886	0.894	0.886	0.896	0.985	0.985

The Overall Novel Infrastructure


Company Intranet
Disk Files
Service Data
CRMs
Open Data
ETL
Enterprise-Wide Data Warehouse
Relational Schema
Data Lake
Multidimensional Data Marts (Analysis Layer)
DM1
DM2
...
DMn
Knowledge Extraction & Discovery
Simulation
Presentation & Reporting Layer, User Interface
Data Scientist
Call center agent
Management
Domain Expert


ETL
Decision support infrastructure
SLA,
KPI,
Statistics
Action
RULES
Production systems
Decision layer

References

A. Brunello, P. Gallo, E. Marzano, A. Montanari, N. Vitacolonna, *An event-based data warehouse to support decisions in multi-channel, multi-service contact centers*, 2019.

A. Brunello, E. Marzano, A. Montanari, G. Sciavicco, *A combined approach to the analysis of speech conversations in a contact center domain*.