Venice Airport:
A small Big Data story

Fabrizio Montino
Dario Cannone
Venice Airport in Numbers

9.6 MILLION PASSENGERS

9 LONG HAUL DESTINATIONS

6 NORTH AMERICA

3 MIDDLE EAST

50/100 OVER 50 CARRIERS OVER 100 DESTINATIONS

1 ITALIAN INTERCONTINENTAL OF 3 GATEWAYS

2 HOME BASED CARRIERS VOLOTEA & EASYJET

3 ITALIAN AIRPORT SYSTEM TOGETHER WITH TREVISO APT

5 NORTH AMERICAN INTERCONTINENTAL CARRIERS SERVE VCE

15 MAJOR EUROPEAN HUBS LINKED WITH MULTIPLE DAILY FREQUENCIES

300M € INFRASTRUCTURE INVESTMENTS IN THE LAST 4 YEARS

600M € IN THE COMING 5 YEARS

86% INTERNATIONAL TRAFFIC

29% TRAVEL FOR BUSINESS

46% POINT OF SALE ITALY

26% OF ALL PAX CONNECT TO REACH THEIR FINAL DESTINATION
Venice Airport Catchment Area
A well balanced mix of incoming/outgoing leisure, business traffic and ethnic flows

1. A 8 million people extended-catchment area combined with the largest mediterranean home-port for cruise ships

2. Two of the “must see destinations of the world”, Venice and the Dolomites Mountains, both Unesco World Heritage Sites

3. A large and diversified foreign community, strong driver of ethnic traffic

4. One of the strongest economy in Europe, with a wide manufacturing footprint

INCOMING 65%
- LEISURE: 29%
- BUSINESS: 57%
- ETHNIC/VFRs: 14%

OUTGOING 35%
- LEISURE
- BUSINESS
- ETHNIC/VFRs

VCE: 9.6M pax in 2016 +10% vs PY
Passenger Experience

- Queues, waiting time
- Waste of time
- Worries about losing the flight
- Worries about being late

- Seamless travel operations
- More free time
- Time to relax, work, do shopping, ...
Passenger Experience

- Queues, waiting time
- Waste of time
- Worries about losing the flight
  Worries about being late

- Seamless travel operations
- More free time
- Time to relax, work, do shopping, ...
Airport Resource Management

- Runway
- Aircraft park
- Bus/Boarding bridge
- Baggage handling

When departing...
- Car Parks
- Check-in desks
- Security Checks
- Bars, restaurants, toilets, shops, wi-fi, ...
- Custom controls
- Boarding gates
Airport Resource Management

- Runway
- Aircraft park
- Bus/Boarding bridge
- Baggage handling

When departing...
- **Car Parks**
- Check-in desks
- Security Checks
- Bars, restaurants, toilets, shops, wi-fi, ...
- Custom controls
- Boarding gates
Predicting Car Park Occupation
Predicting Car Park Occupation
Predicting Car Park Occupation
Predicting Car Park Occupation

Optimizing occupation of car parks is not an easy task for an airport, especially for the dimensions of the problem...

- 18 car parks
- around 16000 cars every day!

Being able to predict occupation would be a major advantage!

Our project:
**Predict daily occupation peak**
Some data about the data

- 18 car parks
- almost 6 millions cars every year
- data since late 2009

1 transaction = 2 movements (in and out)

50 millions transactions

100 millions records to analyze...!

This is a small big data problem...
## What can we do?

<table>
<thead>
<tr>
<th>Possible solution</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Load and process all data in KNIME | • The loading part is straightforward  
• Many many knime nodes for transformations | • Processing tables with many million rows can be terribly long |
### What can we do?

<table>
<thead>
<tr>
<th>Possible solution</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Preprocessing in the database, then load data in KNIME | ● Only aggregated into KNIME  
● KNIME Database nodes and SQL for data cleaning | ● DWH becomes overloaded |

**Pros**
- Only aggregated into KNIME
- KNIME Database nodes and SQL for data cleaning

**Cons**
- DWH becomes overloaded
**What can we do?**

<table>
<thead>
<tr>
<th>Possible solution</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move everything to a Hadoop cluster</td>
<td>• Performance</td>
<td>Can be very time-consuming (buy servers, install Hadoop, learn Hadoop tools, migrate data, write code…)</td>
</tr>
<tr>
<td></td>
<td>• Scalability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Many useful tools</td>
<td></td>
</tr>
</tbody>
</table>
What can we do?

Think big, but start small

- Yes, let's use Hadoop, but:
  - Let's start with a minimal cluster,
  - Let's focus on just one use case to develop on Hadoop...
  - ...and use KNIME to exploit Hadoop tools!

This is where KNIME showed us how helpful it can be
Data Loading

We want to move the table of parking transactions to Hadoop
Data Cleaning

All transformations are done in Hadoop!

- Date and string manipulations
- Joins
- Missing Values
From Transactions to Occupancy
Enriching the dataset

Besides year, month and day, we can add many more features:

- Weekends and holidays
- **Distance to nearest holiday**
- Average occupation of previous years
- **Departing or arriving flights**
- Sum of passengers
Data Mining, at last

We finally import an **aggregated, cleaned, enriched dataset** into KNIME.

For each parking area, we train:

- Lasso/Ridge Regressions
- Gradient Boosted Trees
Data Mining, at last

Cross-Validation

- Training
- Test

From DWH to Hadoop
Preprocessing and cleaning
Building the model
Deployment
Data Mining, at last

**Cross-Validation**

- Training
- Test

**Forward Chaining**

- Training
- Test

**Flowchart**

1. From DWH to Hadoop
2. Preprocessing and cleaning
3. Building the model
4. Deployment
Data Mining, at last
Some Results

Predicted Occupation
Park 0

Park 0
Prediction error (moving average)

From DWH to Hadoop  Preprocessing and cleaning  Building the model  Deployment

Legend
- Predicted Values
- Estimated Errors

MA(relative_error)
### Publishing results with KNIME Server

#### KNIME WebPortal

<table>
<thead>
<tr>
<th>Car Park</th>
<th>Start date</th>
<th>End date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Date: 2017-01-01</td>
<td>Date: 2017-03-01</td>
</tr>
</tbody>
</table>

#### Data Table

<table>
<thead>
<tr>
<th>park</th>
<th>date and time</th>
<th>weekday</th>
<th>holidays</th>
<th>season</th>
<th>distance to holiday</th>
<th>pax_d</th>
<th>pax_a</th>
<th>not_schengen</th>
<th>day_peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2017-02-20</td>
<td>Monday</td>
<td>None</td>
<td>Winter</td>
<td>45</td>
<td>15012</td>
<td>14770</td>
<td>0.272</td>
<td>1463</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-21</td>
<td>Tuesday</td>
<td>None</td>
<td>Winter</td>
<td>45</td>
<td>11914</td>
<td>12114</td>
<td>0.254</td>
<td>1487</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-22</td>
<td>Wednesday</td>
<td>None</td>
<td>Winter</td>
<td>47</td>
<td>12186</td>
<td>12048</td>
<td>0.262</td>
<td>1567</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-23</td>
<td>Thursday</td>
<td>None</td>
<td>Winter</td>
<td>48</td>
<td>12709</td>
<td>12720</td>
<td>0.274</td>
<td>1532</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-24</td>
<td></td>
<td>None</td>
<td>Winter</td>
<td>49</td>
<td>15637</td>
<td>16246</td>
<td>0.275</td>
<td>1539</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-25</td>
<td></td>
<td>None</td>
<td>Winter</td>
<td>50</td>
<td>13100</td>
<td>12905</td>
<td>0.305</td>
<td>1165</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-26</td>
<td></td>
<td>None</td>
<td>Winter</td>
<td>49</td>
<td>15590</td>
<td>15790</td>
<td>0.27</td>
<td>1156</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-27</td>
<td></td>
<td>None</td>
<td>Winter</td>
<td>48</td>
<td>15240</td>
<td>14848</td>
<td>0.249</td>
<td>1482</td>
</tr>
<tr>
<td>0</td>
<td>2017-02-28</td>
<td></td>
<td>None</td>
<td>Winter</td>
<td>47</td>
<td>12389</td>
<td>12181</td>
<td>0.271</td>
<td>1505</td>
</tr>
</tbody>
</table>
Conclusions

- Venice Airport now has predictive analyses to optimize parking occupation!

- But this is just the beginning:
  - We have now an Hadoop cluster that is ready to scale and to be used for other projects,
  - a KNIME Server to automatize execution and publish results for the businessmen
  - and many ideas to realize!