Data Science in the Travel Industry

Real-world experience using current leading frameworks

Paul Balm
September 2015
Hello!

I am from Amsterdam, since 2005 in Madrid

1999 – 2004: Ph.D. in particle physics at Fermilab in Chicago

2005 – 2014: Data Processing projects for the European Space Agency

2014 – present: Data Scientist at Amadeus Travel Intelligence
Agenda

_ Travel Industry and Amadeus

_ A User Story

_ Experiences with Different Technologies

_ Conclusions & Final Remarks

All views expressed in this presentation are entirely my own
The Travel Industry and Amadeus
Amadeus in a few words

Amadeus is a technology company dedicated to the global travel industry. We are present in 195 countries with a worldwide team of more than 12,000 people. Our solutions help improve the business performance of travel agencies, corporations, airlines, airports, hotels, railways and more.
We started our journey in 1987
A history of shaping the future of travel

1987
Amadeus is founded

1992
First booking made through Amadeus

1995
World leader in number of travel agency locations

1996
e-Commerce division launched

1998
1 MILLION bookings made on a single day for the first time

2000
Partnership with BA and Qantas to launch Amadeus Altéa – our core Airline IT offering

2010
Amadeus diversifies into IT for hotel, rail and airport

2014
Contracts with Ryanair and Southwest Airlines and a strategic technology relationship with IHG
Robust global operations

- 1.6+ billion data requests processed per day
- 525+ million travel agency bookings processed in 2014
- 695+ million Passengers Boarded (P Bs) in 2014
- 95% of the world's scheduled network airline seats
Connecting the entire travel ecosystem

- Cruiselines
- Hotels
- Car rental
- Ground handlers
- Ferry operators
- Ground transportation
- Airports
- Travel agencies
- Insurance companies
- Airlines
Supporting the entire traveller life cycle

- Post-trip
- Inspire
- On trip
- Search
- Pre-trip
- Buy/Purchase
A User Story
A Network Carrier

An airline operating a network with hubs

Any risks or opportunities for this operation?
Schedule Analysis

Analyse global schedules to alert of any changes in capacity

Schedule information may come from
  _Commercially provided feeds
  _Airline inventory data
  _Publicly available databases (EuroStats, US DoT)
Schedule Analysis

Capacity share: Amsterdam – Berlin

Sources: OAG schedules
Schedule Analysis

Batch processing of different formats of input data

_Different formats, different sources and a priori unclear which sources will be useful or when_

_Store everything always_ – within legal and moral boundaries

_Central data site with large NAS (network attached storage)_

_HDFS based storage at hosted processing clusters_

_Maintain separation of different customer projects_
A Network Carrier

An airline operating a network with hubs

Schedule analysis shows competitor entering the market

What’s the risk?
Quality of Service

QSI: Quality of Service Index
The prime quality metric: Time to get from A to B
In practice many rules to build QSI that require tuning and iteration

Agility as a requirement

Need for scalable processing capacity
A Network Carrier

- An airline operating a network with hubs

- Schedule analysis shows competitor entering the market

- Risk of loss of market share assessed via QSI analysis

- How to react?
Future revenue streams

Information about future revenue comes in many forms:
- Tickets issued
- Reservations (foreseen revenue)
- Searches
- Capacity changes
- Sentiments on social networks, weather forecasts and the like

- How are average fares developing?
- How are passenger counts developing?

Requires fast updates – live if possible
Streaming?

No hard truths here – considerations of options only!

- Response times of the order of seconds are not required in the presented use-case
- We still require to process the large amount of historical data
- We think processing the historical data is better done in batch
  - Mature and reliable approach
  - Can we reuse the streaming code for batch processing?

- Flink: Suitability for batch processing to be demonstrated
- Storm: Using Summingbird, combine with Scalding?
- Spark Streaming: Combine with Spark for batches. Maturity?
A Network Carrier

An airline operating a network with hubs

Schedule analysis shows competitor entering the market

Risk of loss of market share assessed via QSI analysis

How to react?
Technologically
The Travel Intelligence Engine

Scalable cloud-based Hadoop platform

- Amadeus Data (Altéa and other)
- Customer Private Data
- Industry
- Weather
- Demographics
- Social Media
- Etc.

Raw Data Storage

Batch and streaming processing

Output storage

Visualization layer

Applications

Data Warehouse

(* All the raw data is kept after processing to allow for re-processing when required)
Lambda architecture

Image reproduced from http://lambda-architecture.net/
Scoobi

batch layer

master dataset

new data

speed layer

real-time view

real-time view

serving layer

batch view

batch view

query

query
Benchmarking Spark vs. Scalding by @BoxHQ

Source: Box Engineering Blog
Web Services

Visualization tools
Web Services
- Python: Flask
- Scala: Spray

Visualization tools

Scoobi
Scalding
Kafka
Storm
Impala
Web Services
Python: Flask
Scala: Spray
Web Services
- Python: Flask
- Scala: Spray

Visualization Tools
- MicroStrategy
- Tableau
- QlikView

HTML5/JS
Recap

Data science opportunities in the travel sector

How we are dealing with these opportunities at Amadeus
Travel Intelligence

An overview of our experience with different tools

What’s on our horizon for evaluation and why
Some final remarks

Different technologies currently under evaluation

In all areas (batch, streaming, visualizations) too many options to carefully evaluate all
Also, development effort from the community is diluted

Consolidation would be welcome!

Even so, tools are advancing rapidly and within their limitations, many are of great value
Thank you

@paul_balm
https://es.linkedin.com/in/paulbalm