Data Science and Big Data in Travel Industry

Amadeus Travel Intelligence Use Cases

Anwar Rizal
Emmanuel Bastien

Outline

Amadeus Travel Intelligence

Data Analytic Tasks

Use Cases

- e-Commerce Conversion Rate
- Airline Customers Segmentation

Technology Point of View

Summary and Conclusion

amadeus

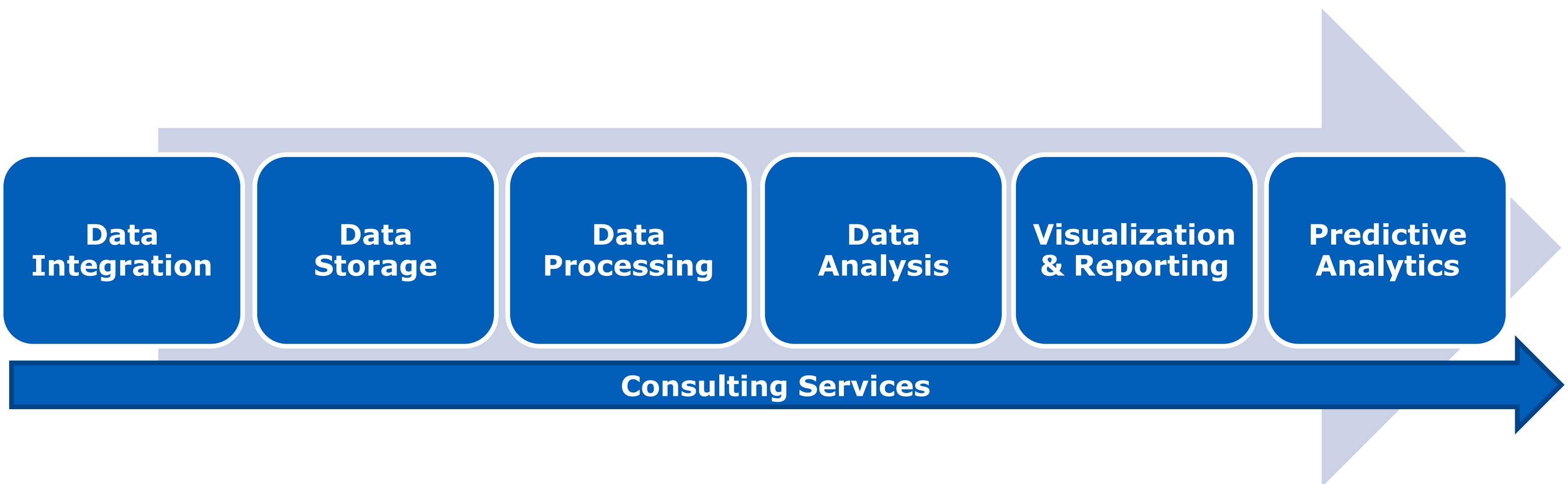
1 Amadeus Travel Intelligence

Our mission:

"To provide **unique** *and* **actionable insights** *to each of our customers using advanced technologies"*

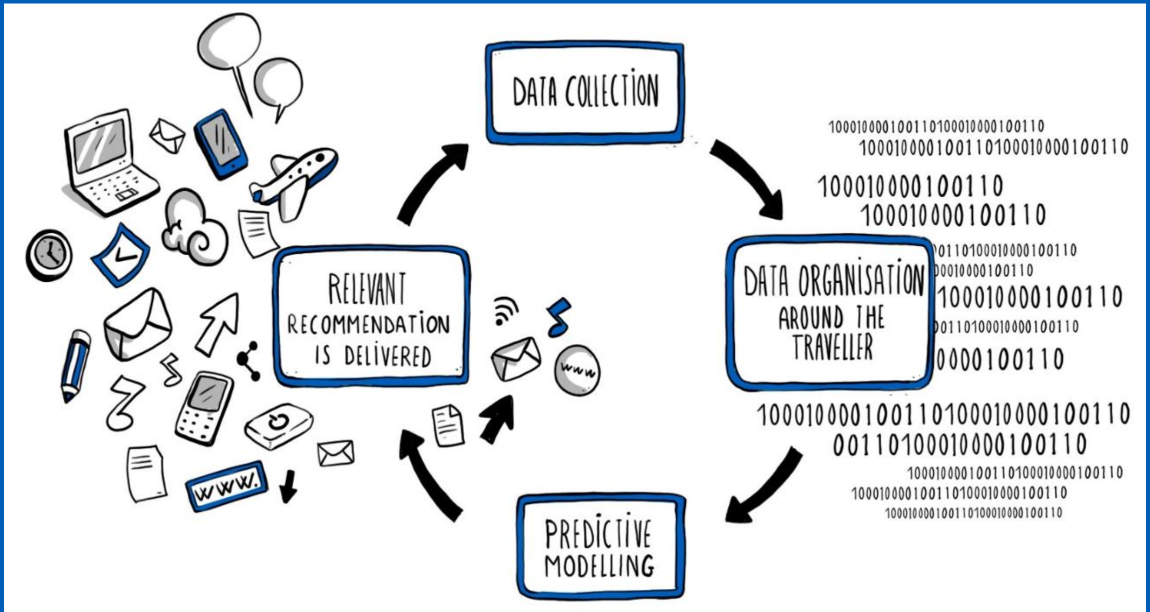
Our Customer Segments

Our Customer Segments & Value Chains



2

Data Analytic Tasks



Starting Points

_ Start from business requirements

- “I want to act when the number of booking for a given origin and destination decreases”
- “I want to personalize marketing campaigns to my customers”

_ Do not start from data

- “I have all the logs that record everything. I know it’s valuable. What can we do with them ? “

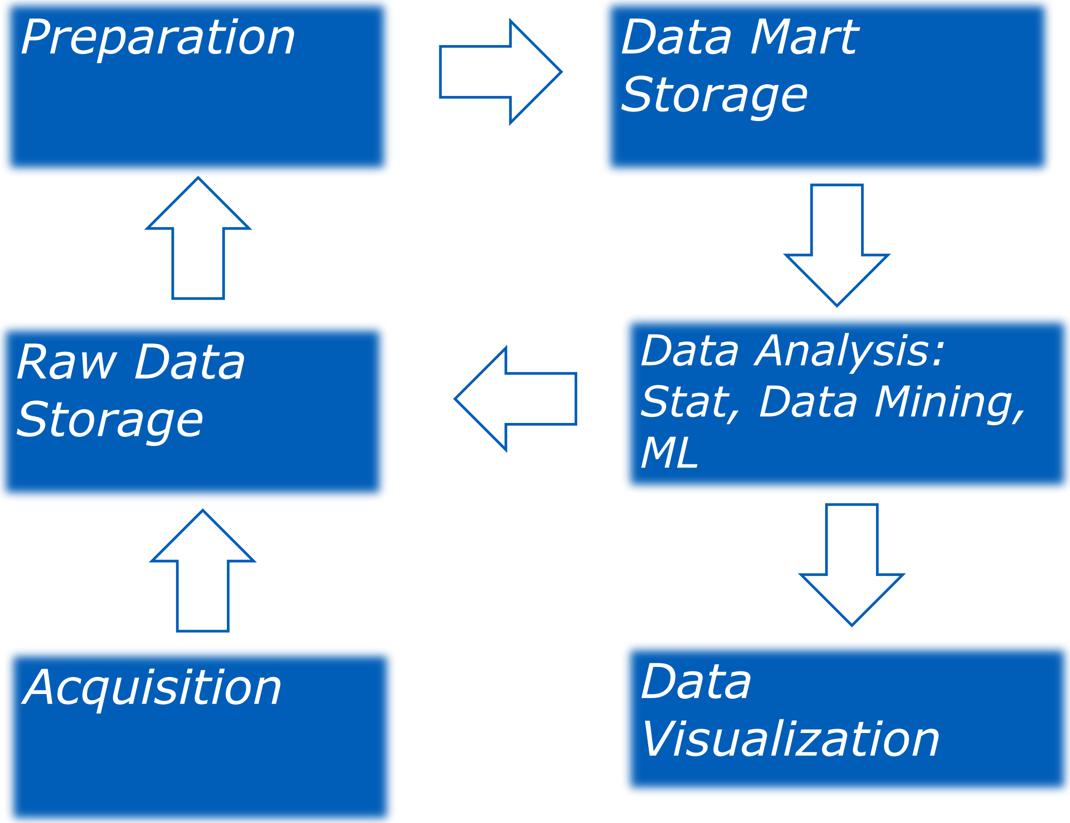
_ Do not start from data analysis activity

- “I want to cluster my passengers”
- “I want to apply machine learning to my data”

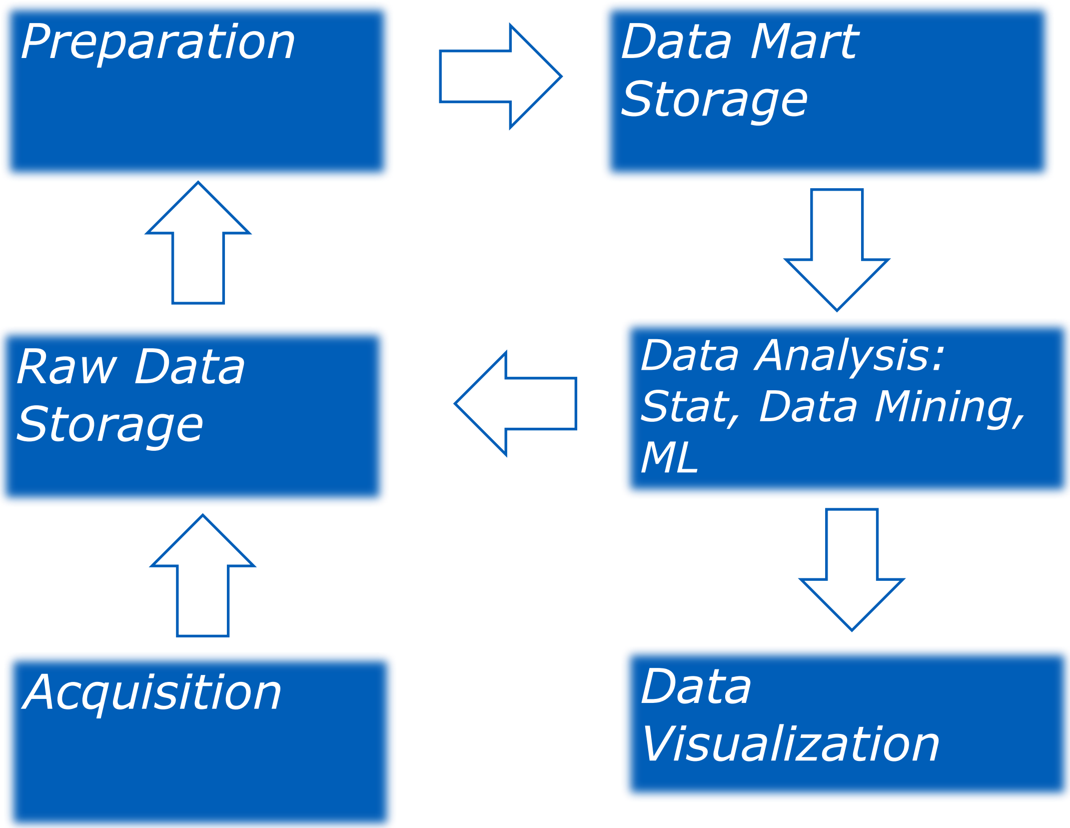
_ Do not start from technology

- “I know we can solve our business problem using Hadoop”

Data Analysis Workflow

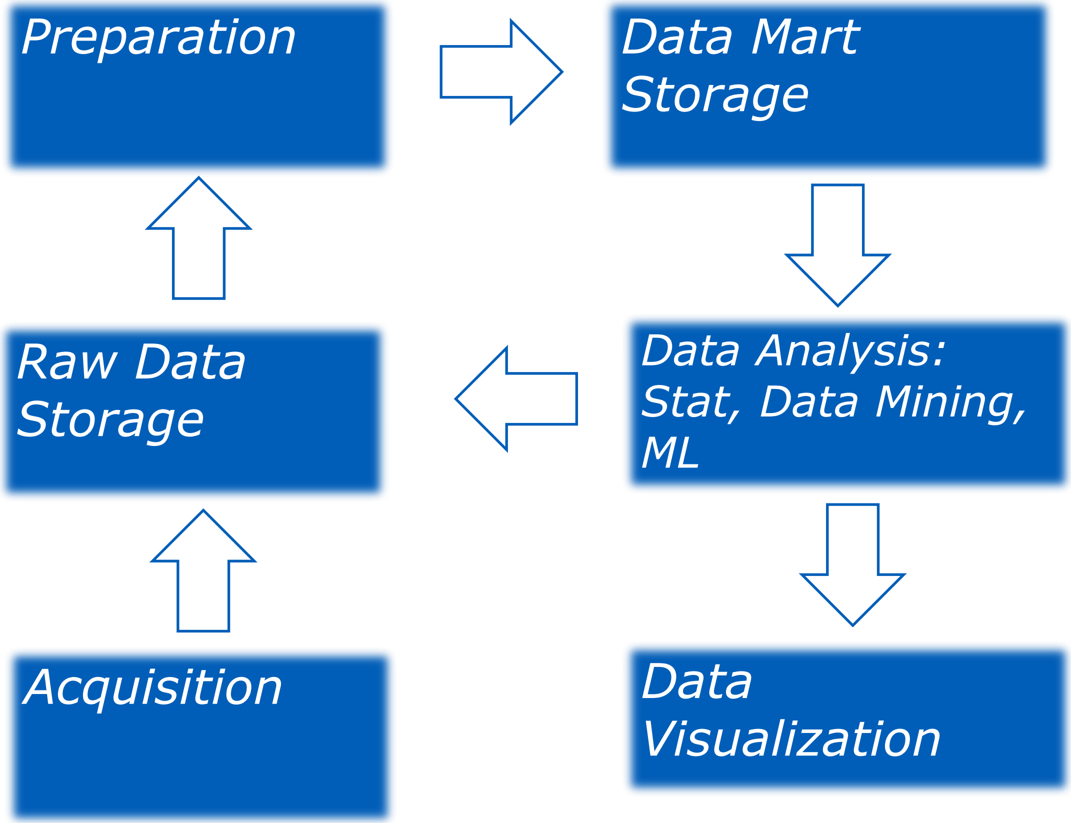


Data Analysis Workflow



log files, Amadeus feeds, external feeds, messaging system, web scrapping, ftp, rsync, …

Data Analysis Workflow

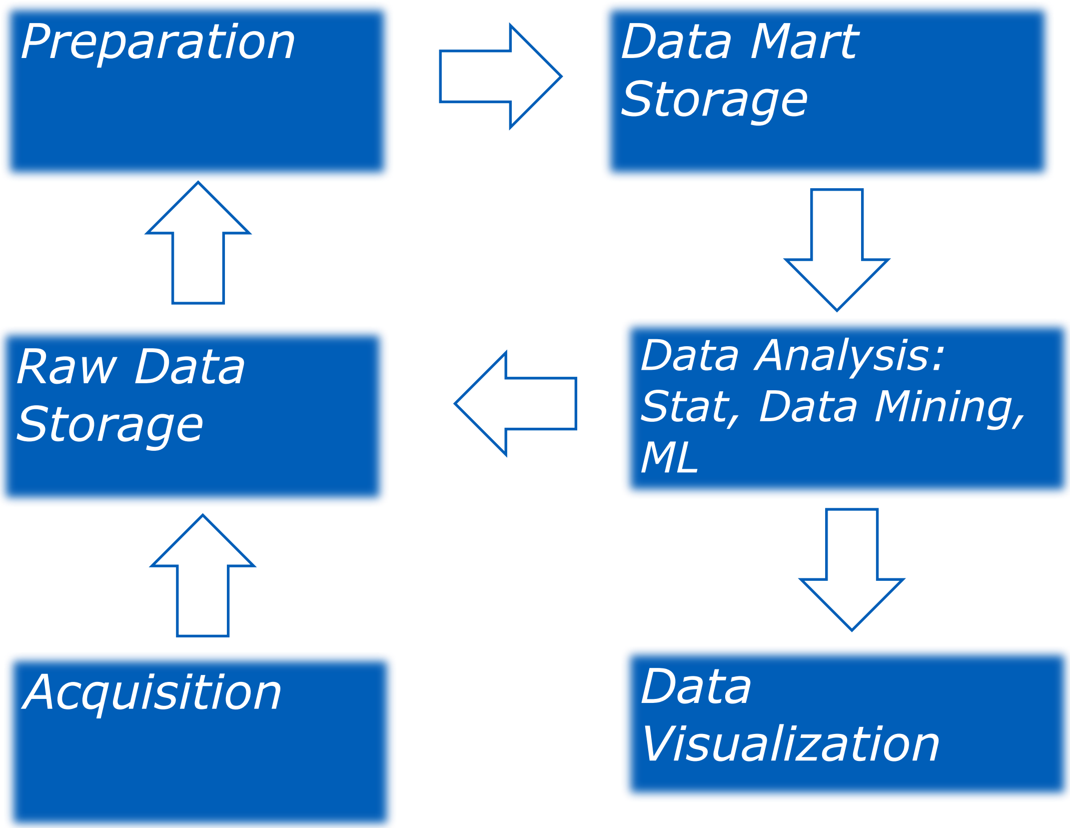


Buch of disks, distributed file system, HDFS, compression, checksum,

amadeus

Data Analysis Workflow

shell commands, visualization for exploration, Hadoop, Spark, ETL



DATA INTEGRATION **DATA PROCESSING** **DATA PUBLICATION** **VISUALISATION /AUTOMATION**

amadeus

The Importance of Data Preparation

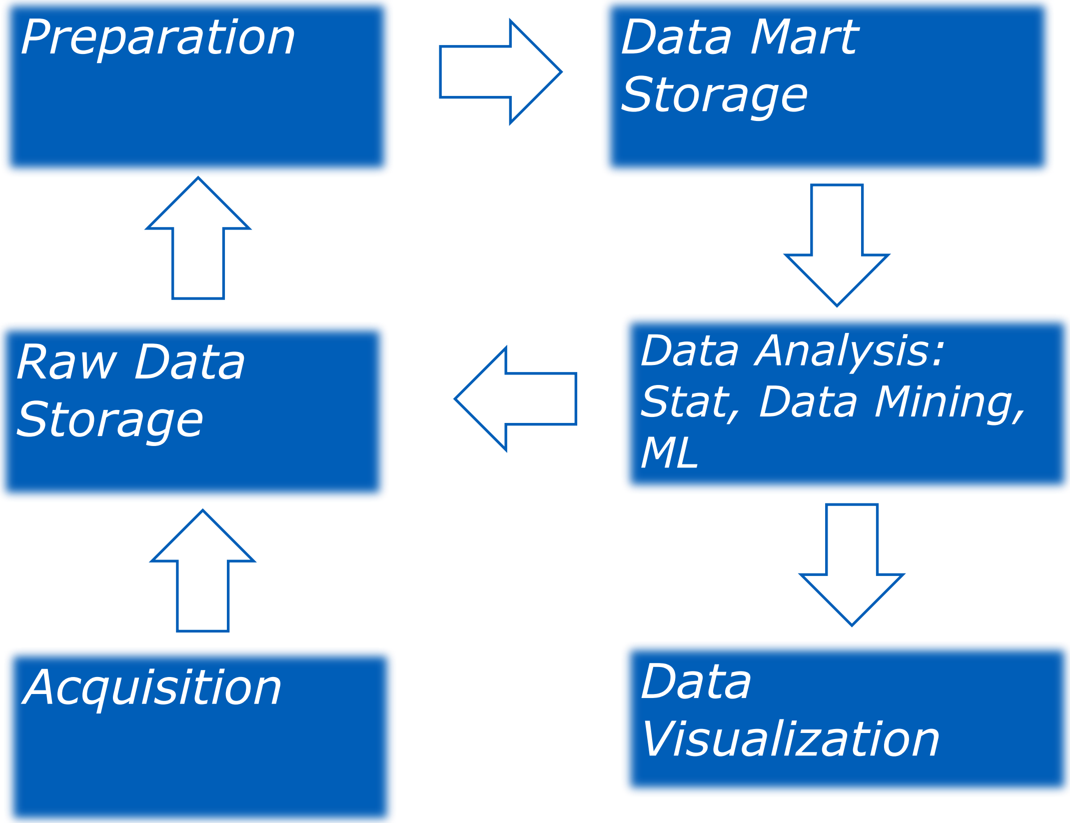
_ The real data:
- Are Incomplete
- Are Buggy
- Come from different sources, and the quality might vary depending on those sources

_ 80 % of data analysis efforts are on data preparation (exploration, cleansing, normalization, data imputation, …)

_ Understanding the quality of the input is important in estimating confidence of the analysis result

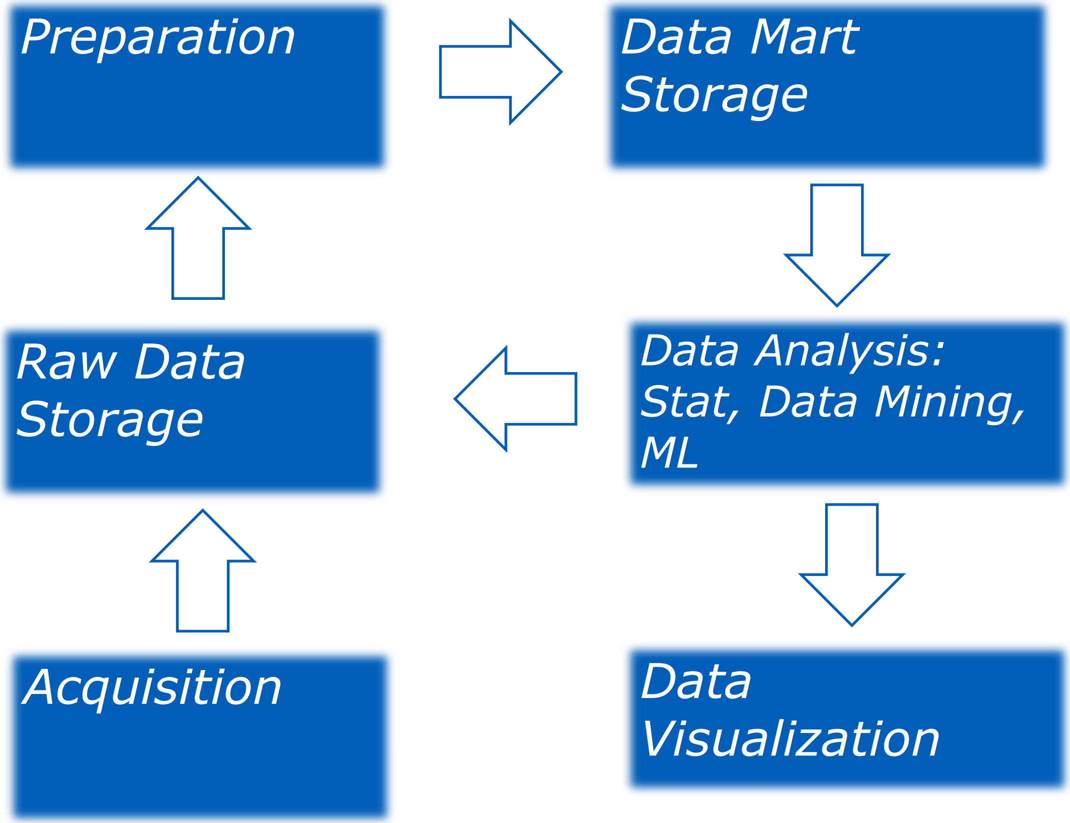
amadeus

Data Analysis Workflow



Data model, OLAP, SQL, No SQL, Key value, Query Engines, …

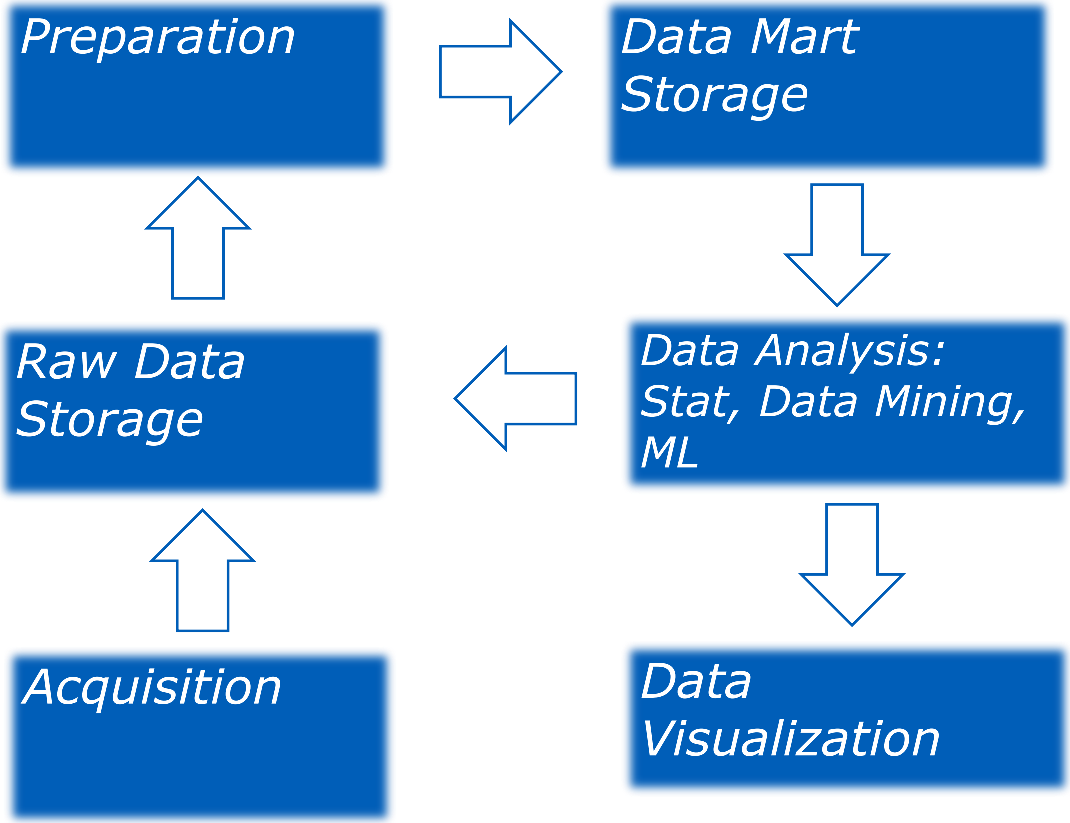
Data Analysis Workflow



Counting, summarizing, modeling, data mining, statistical machine learning,

Hadoop, Spark

Data Analysis Workflow



HTML5, java script, Tableau, Qlik, …

Data Analysis Workflow

shell commands, visualization for exploration, Hadoop, Spark, ETL

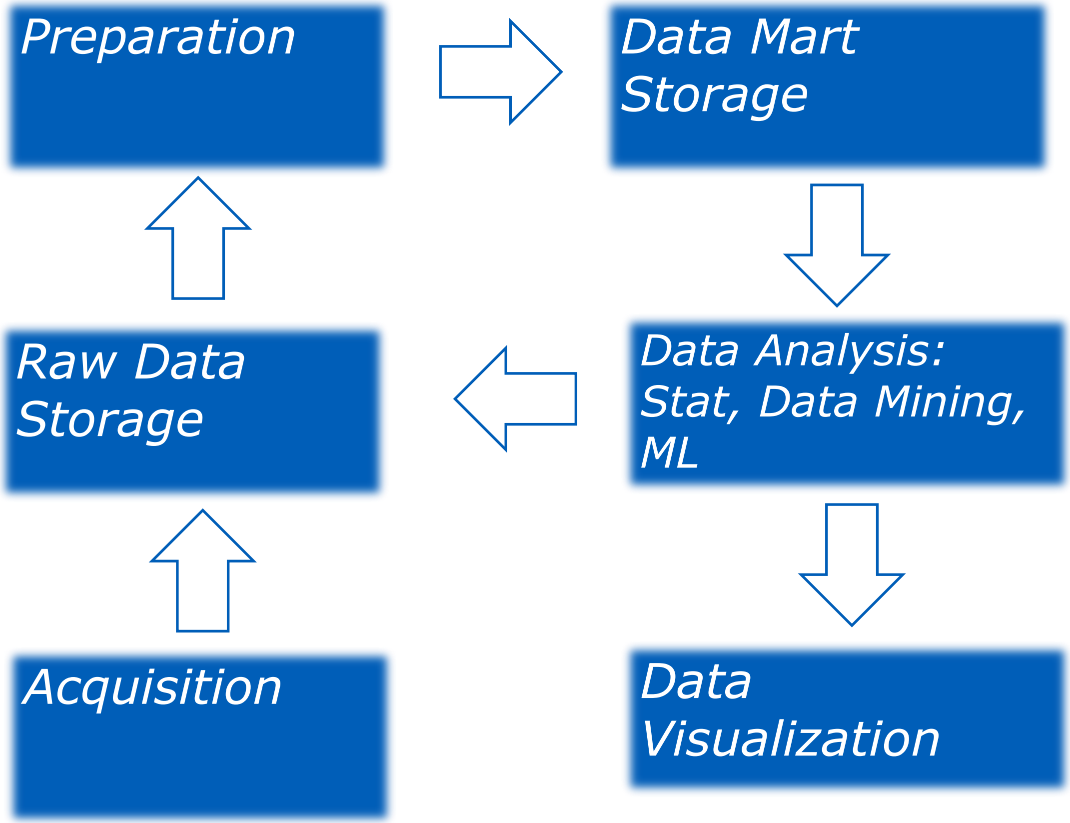
Data model, OLAP, SQL, No SQL, Key value, Query Engines, …

Buch of disks, distributed file system, HDFS, compression, checksum,

Counting, summarizing, modeling, data mining, statistical machine learning

log files, Amadeus feeds, external feeds, messaging system, web scrapping, ftp, rsync, …

HTML5, java script, Tableau, Qlik, …



Data Analysis Important Skills

- Understanding of business requirements and audience
  - To identify the data required to answer the business questions
  - To evaluate and conduct the appropriate data analysis techniques depending on the targeted audience
- Automation of data preparation
- Data analysis using statistics, machine learning, and data mining techniques
- Creation of compelling and meaningful data visualization and telling the story
- Estimating the confidence level to the result of the analysis

Batch vs Real Time

- Traditionally, data analysis are done in batch mode: daily, weekly, monthly, yearly
- Often times, analysis is only possible when the whole data are available for analysis
- Big Data platform such as Hadoop or Spark are powerful tools to do the batch data analysis in large scale

Batch vs Real Time

- More and more companies look for accomplishing business actions based on data in real time
- More and more data are continuously generated, e.g. IOT
- For the time constraint, it is often not possible to process the whole data
- A new set of techniques and algorithms are developed to answer this real time requirement

Batch vs Real Time

- Streaming algorithms get popular to address the real time constraint

- The algorithms make trade-off between the time of execution and precision

- Examples:
  - Bloom Filter
  - Sketch-based Algorithms
  - Hyperloglog
  - Approximate histogram
  - ….
- See *Mining of Massive Datasets* (Leskovec, et.al. 2014) and *Data Streams Models and Algorithms* (Aggarwal, 201)

Batch vs Real Time

- From architectural point of view: Lambda Architecture and Kappa Architecture

- The architectures focus on how to combine the historical data and newer data to answer the user query

- See *Big Data* (Nathan, 2015) and *Questioning the Lambda Architecture* (Kreps, 2014)

3 Use Cases

3.1

Use Case 1: E-Commerce Conversion Rate

Use Case 1: E-Commerce Conversion Rate

_Customer

- Airlines, E-Commerce department

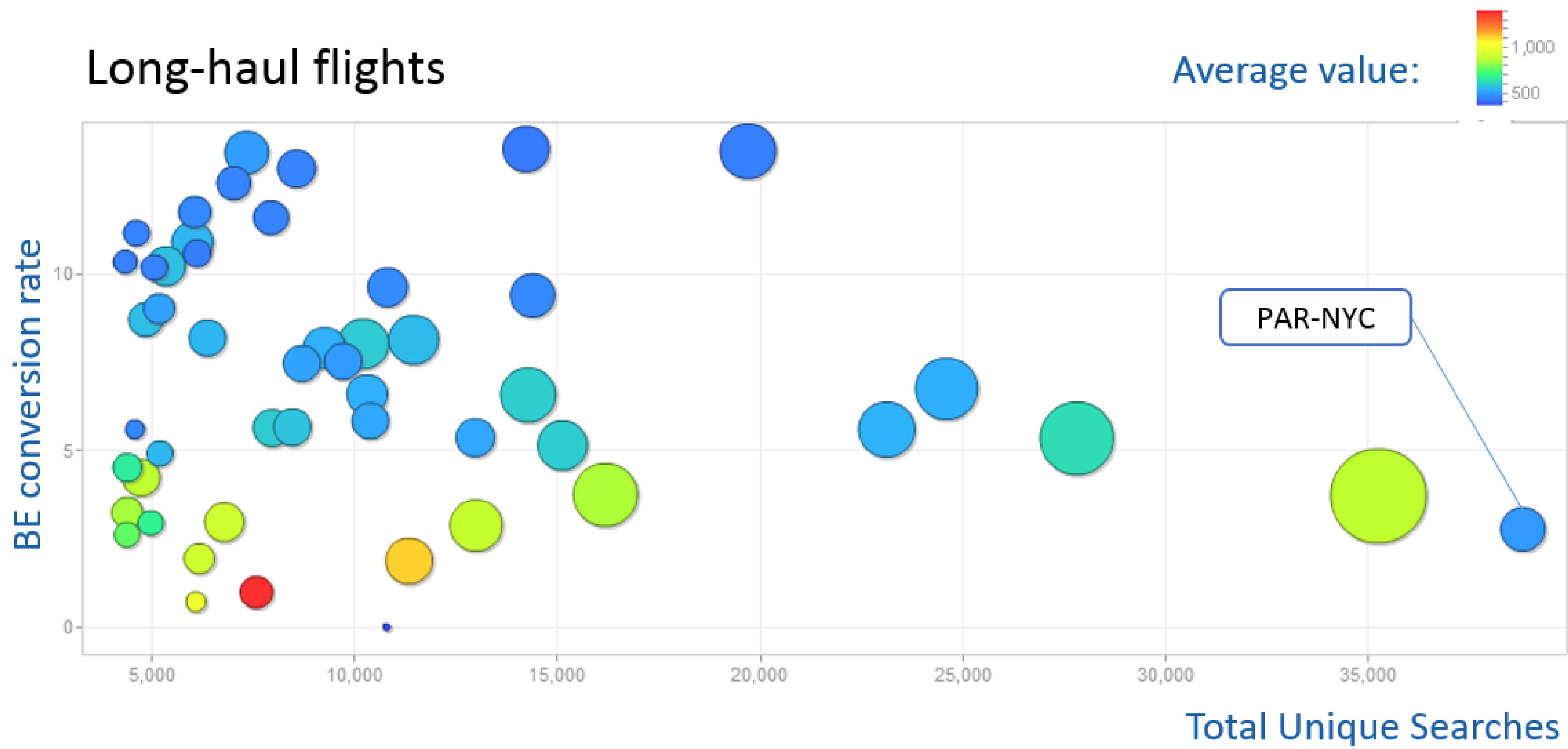
_Business Objective

- Increase the e-commerce revenue

_Business Requirements

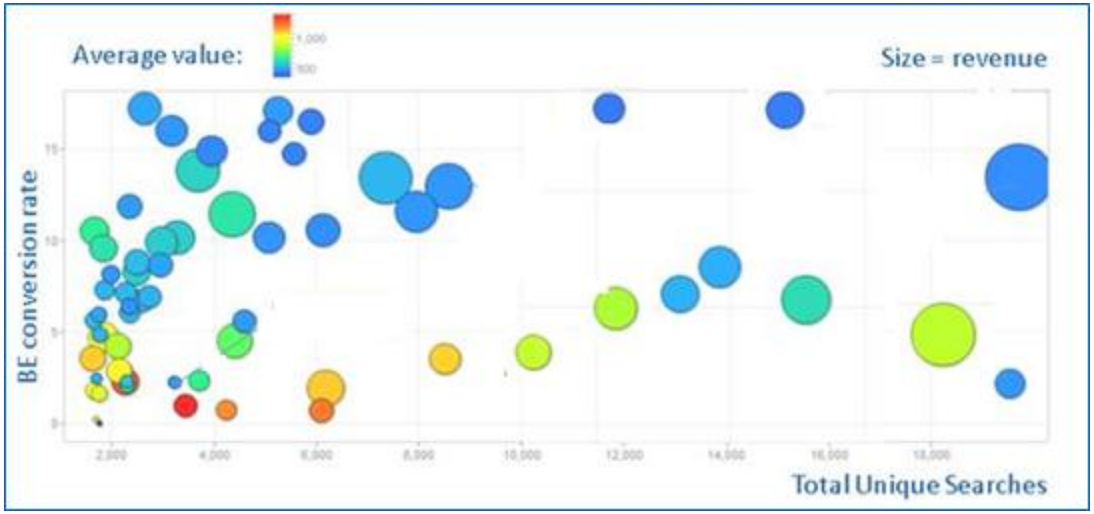
- Increase conversion rate (the ratio of search / booking)
  - Need to have insights on the performance of each product proposed (e.g. Origin and Destination)

Use Case 1: E-Commerce Conversion Rate



The diagram is for Illustration only, it is not derived from real data

Use Case 1: E-Commerce Conversion Rate



The diagram is for Illustration only, it is not derived from real data

amadeus

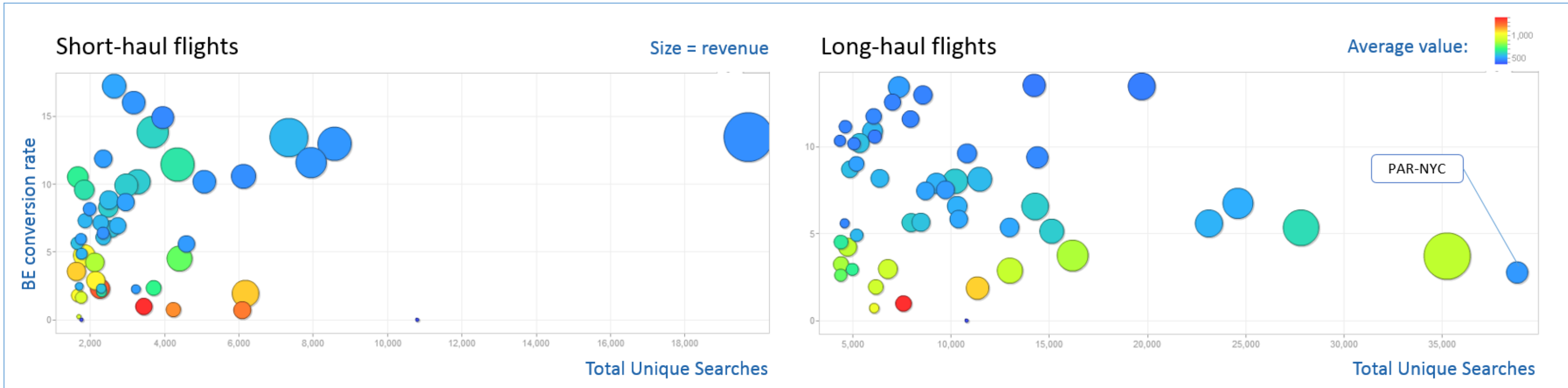
Use Case 1: E-Commerce Conversion Rate

Ori.	Des.	Price position vs. Lowest price in %: Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PAR	NYC	3	1	25	23	27	20	1	5	4	-2	-1	[illegible]
[illegible]	[illegible]	-10	-12	-14	-20	-15	-20	-15	-10	-20	-22	[illegible]	[illegible]
[illegible]	[illegible]	-10	-12	-11	-12	-23	-15	-2	-3	-7	-15	[illegible]	[illegible]
[illegible]	[illegible]	-1	-2	-4	-2	1	2	3	5	10	7	8	[illegible]
[illegible]	[illegible]	0	1	1	2		3	2	1	2	10	-2	[illegible]
[illegible]	[illegible]	3	1	25	23	20	15	13	5	4	-2	-1	[illegible]
[illegible]	[illegible]	-10	-5	-6	-4	-5	-2	-2	-5	-2	-10	-15	-20
[illegible]	[illegible]	20	15	10	4	3	2	-2	-5	-15	-20	-23	-25

The diagram is for Illustration only, it is not derived from real data

amadeus

Use Case 1: E-Commerce Conversion Rate



Ori.	Des.	Type	Market	Nb of Searches	YoY	Avg. PNR value	YoY	Rev.	YoY	Convert. rate	YoY	Nb. Competi.
PAR	NYC	Long	A	38,500	9%	630	-2%	24,255,000	7%	3%	-15%	4
[illegible]	[illegible]	Long	A	35,000	15%	830	-2%	29,050,000	13%	4%	1%	1
[illegible]	[illegible]	Long	A	28,000	-5%	602	-3%	16,856,000	-8%	6%	5%	1
[illegible]	[illegible]	Long	B	24,000	-3%	550	3%	13,200,000	0%	7%	-3%	2
[illegible]	[illegible]	Long	B	23,000	-2%	520	-5%	11,960,000	-7%	6%	-20%	3
[illegible]	[illegible]	Long	B	19,000	3%	560	-6%	10,640,000	-3%	4%	-15%	2
[illegible]	[illegible]	Long	C	17,000	23%	620	5%	10,540,000	28%	5%	2%	3
[illegible]	[illegible]	Long	C	15,000	12%	450	-8%	6,750,000	4%	3%	-2%	4

Load Factor YoY variation in %

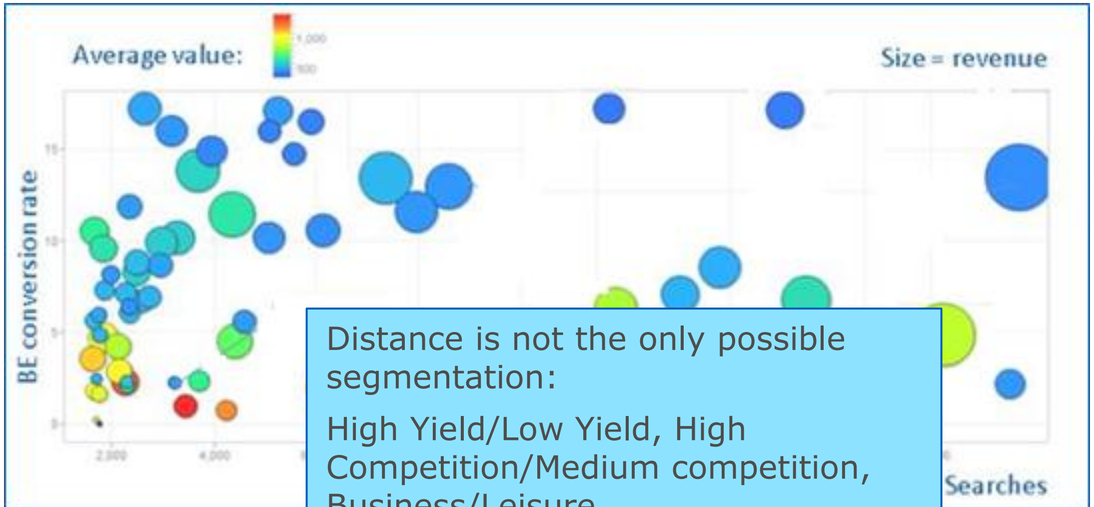
Ori.	Des.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PAR	NYC	5	-6	-6	-10	-17	-15	-18	-20	5	6	10	25
[illegible]	[illegible]	1	2	4	5	2	3	5	6	8	10	2	-1
[illegible]	[illegible]	-1	-2	1	2	3	4	5	15	20	22	23	10
[illegible]	[illegible]	-1	-1	-3	-1	-5	-6	-8	-9	-10	-12	1	10
[illegible]	[illegible]	1	5	10	15	20	25	18	10	1	-10	-5	-9
[illegible]	[illegible]	-10	-10	-15	-17	-15	-15	-18	-20	5	6	10	25
[illegible]	[illegible]	2	3	5	2	1	3	4	5	2	6	8	4
[illegible]	[illegible]	1	2	8	3	25	4	4	5	5	4	2	4

Price position vs. Lowest price in %

Ori.	Des.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PAR	NYC	3	1	25	23	27	20	1	5	4	-2	-1	-15
[illegible]	[illegible]	-10	-12	-14	-20	-15	-20	-15	-10	-20	-22	-25	-30
[illegible]	[illegible]	-10	-12	-11	-12	-23	-15	-2	-3	-7	-15	-25	-30
[illegible]	[illegible]	-1	-2	-4	-2	1	2	3	5	10	7	8	10
[illegible]	[illegible]	0	1	1	2		3	2	1	2	10	-2	-1
[illegible]	[illegible]	3	1	25	23	20	15	13	5	4	-2	-1	-15
[illegible]	[illegible]	-10	-5	-6	-4	-5	-2	-2	-5	-2	-10	-15	-20
[illegible]	[illegible]	20	15	10	4	3	2	-2	-5	-15	-20	-23	-25

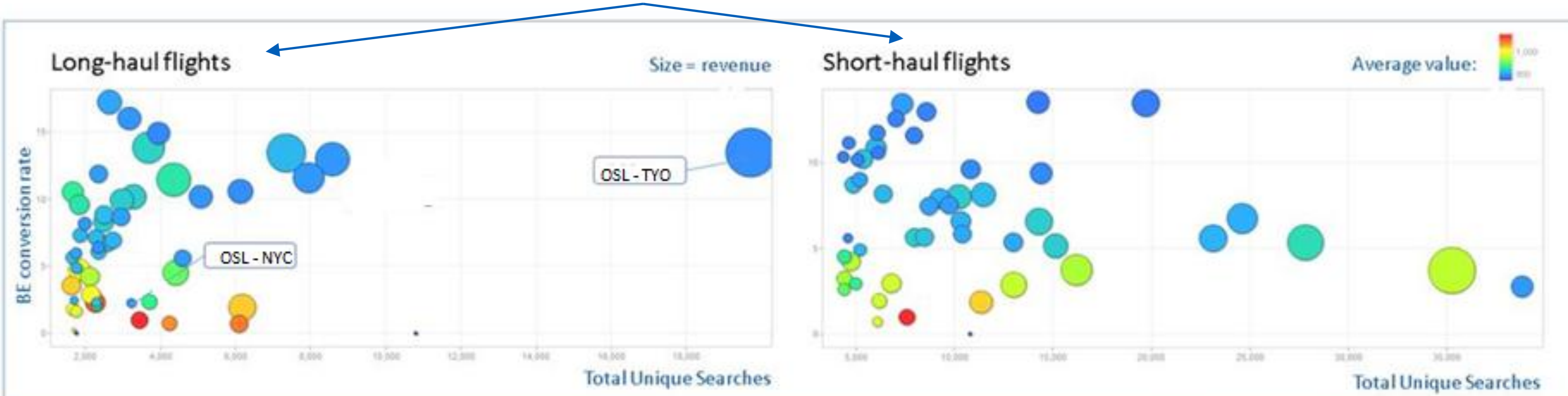
The diagram is for Illustration only, it is not derived from real data

Use Case 1: E-Commerce Conversion Rate



Distance is not the only possible segmentation:

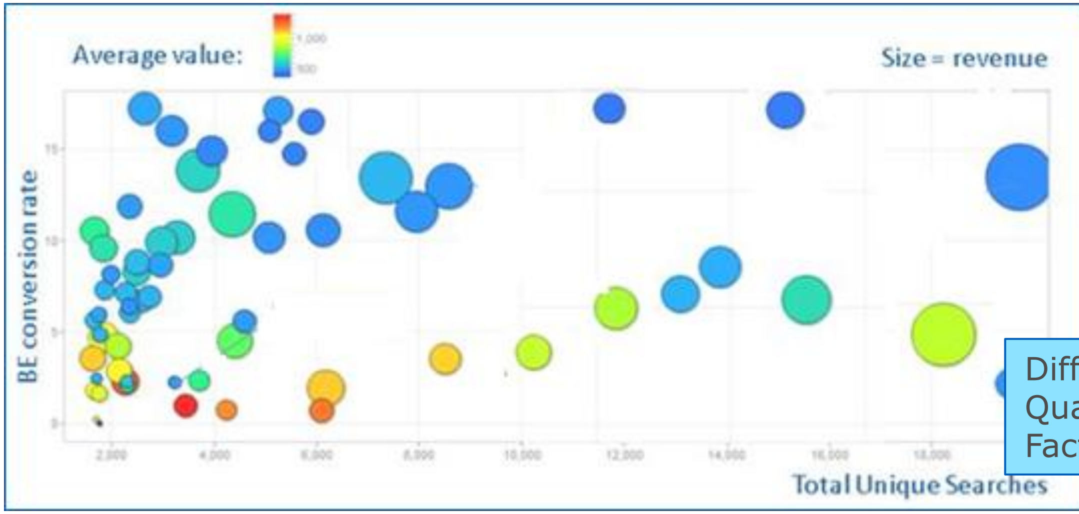
High Yield/Low Yield, High Competition/Medium competition, Business/Leisure, …



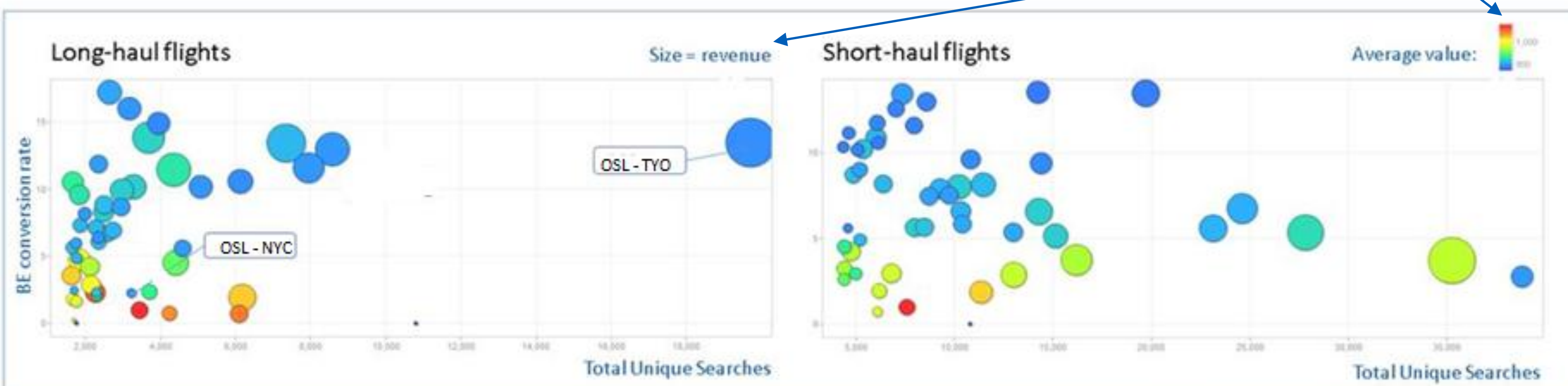
The diagram is for Illustration only, it is not derived from real data

amadeus

Use Case 1: E-Commerce Conversion Rate

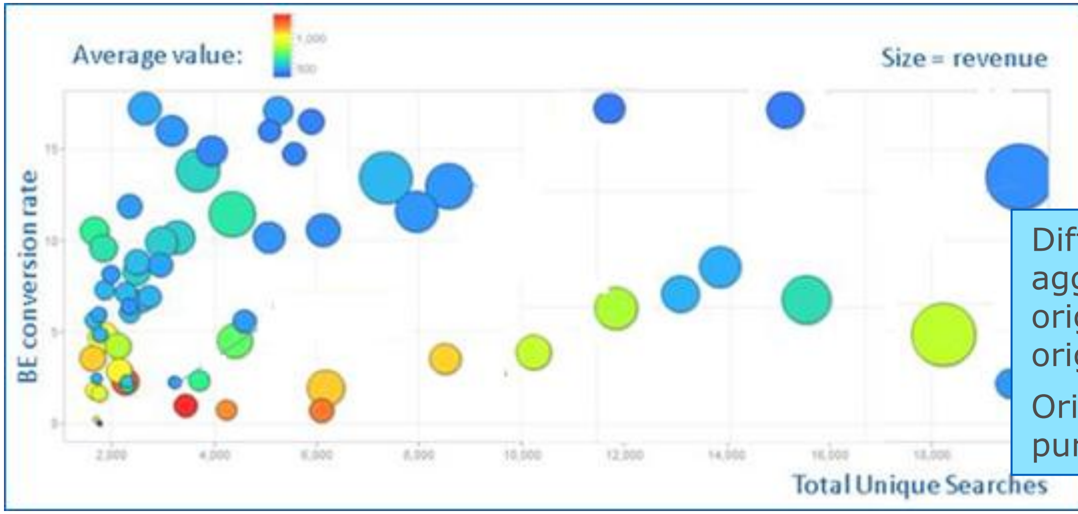


Different features: Yield, Quality of services, RPK, Load Factor, …



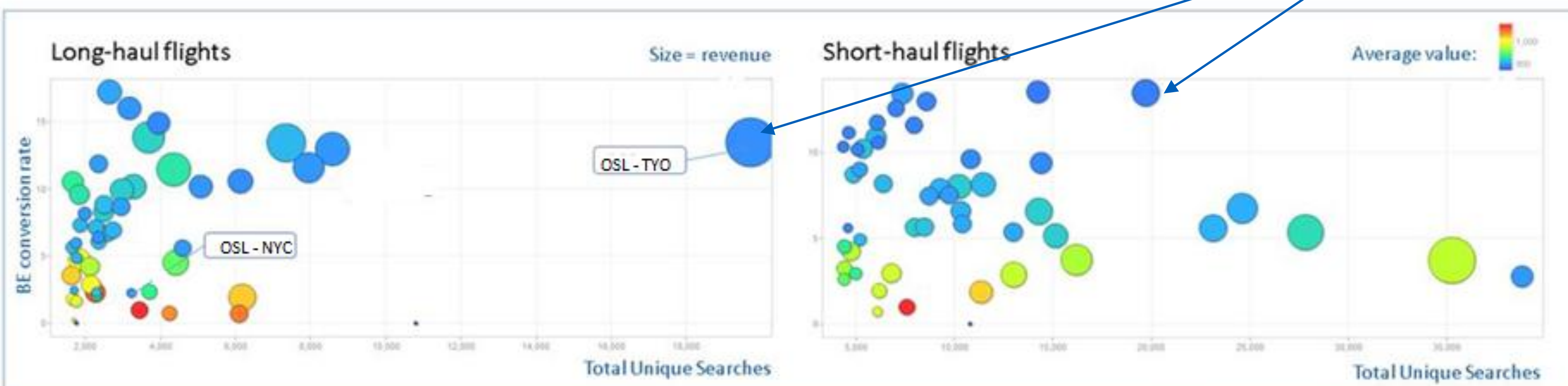
The diagram is for Illustration only, it is not derived from real data

Use Case 1: E-Commerce Conversion Rate



Different aggregation/dimension: origin/destination, origin/destination/type of stay,

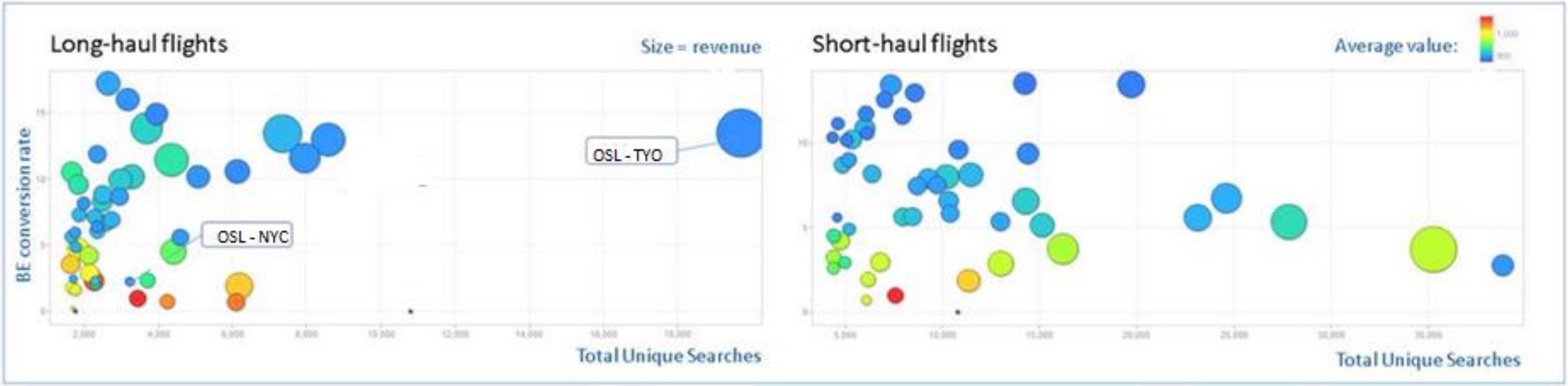
Origin/destination/advance purchase, …



The diagram is for Illustration only, it is not derived from real data

Use Case 1: E-Commerce Conversion Rate

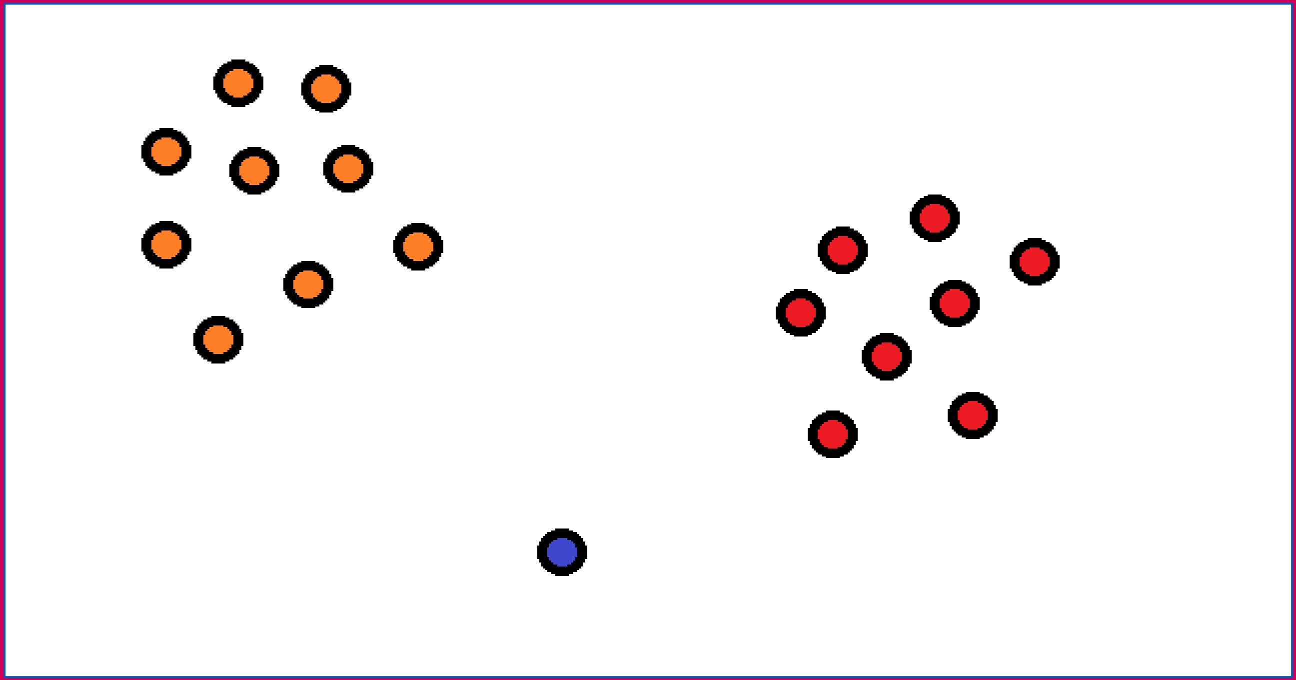
Outlier List					
Origin	Destination	Type of Stay	Advance Purchase	Segmentation	Features
OSL	BGO	Short		Distance	Average Fare
OSL	PAR	Short		High Competition	Average Fare
STO	TYO		Long	Distance	Average Fare



The diagram is for Illustration only, it is not derived from real data

amadeus

3.1.1 Outlier Analysis in 5 Slides



AMADEUS

Outlier Analysis

_ For univariate and data following a normal distribution (or assumed to be so):

- Calculate the probability of the occurrence of such data

  average of search=4500

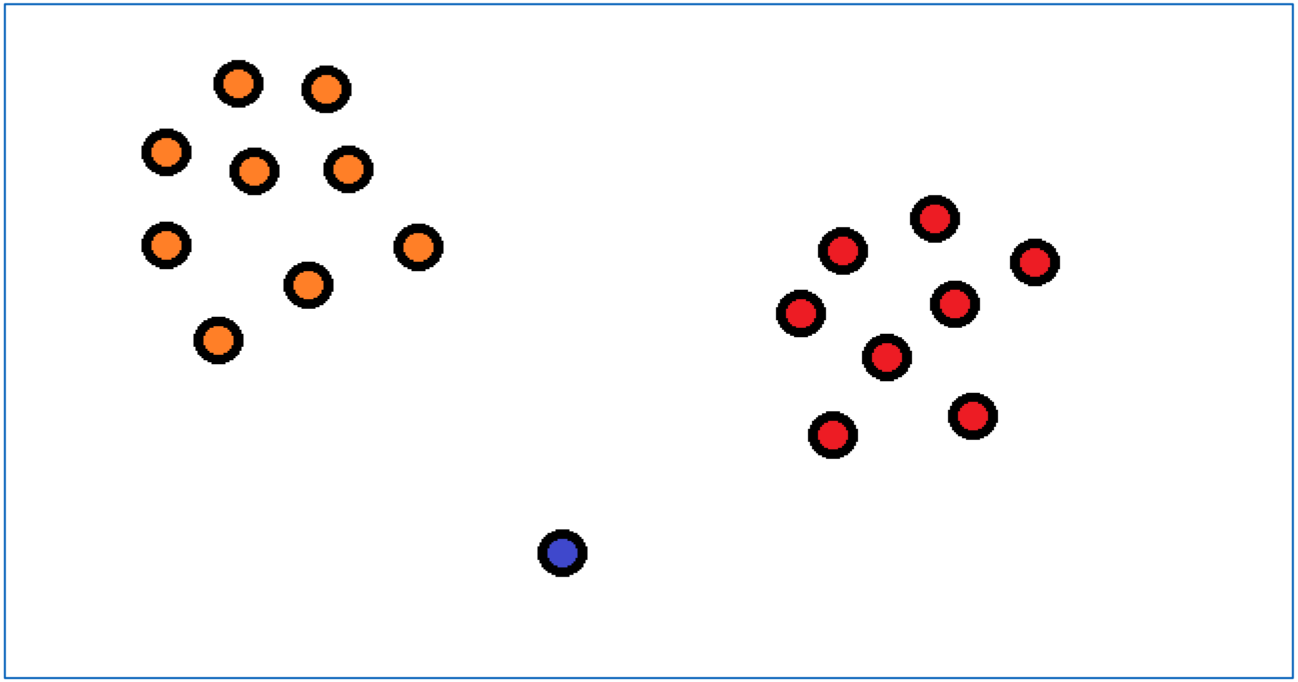
  standard deviation=1000

  point to be checked has search count = 400

  probability = 2.06 x 10-5 => outlier

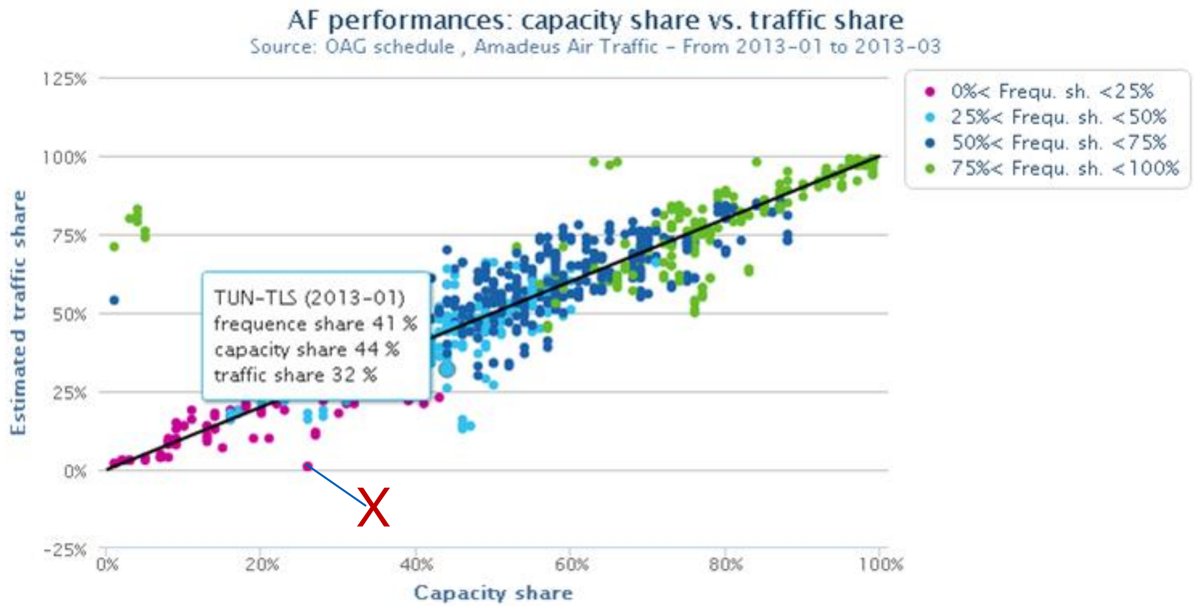
_ We might want to use the average of its group (e.g. average of search for all O&D in blue)

Outlier Analysis



- Blue point is outlier, because:
  - Its distance to the centroid of two other clusters are relatively far, or
  - It is in a cluster of its own

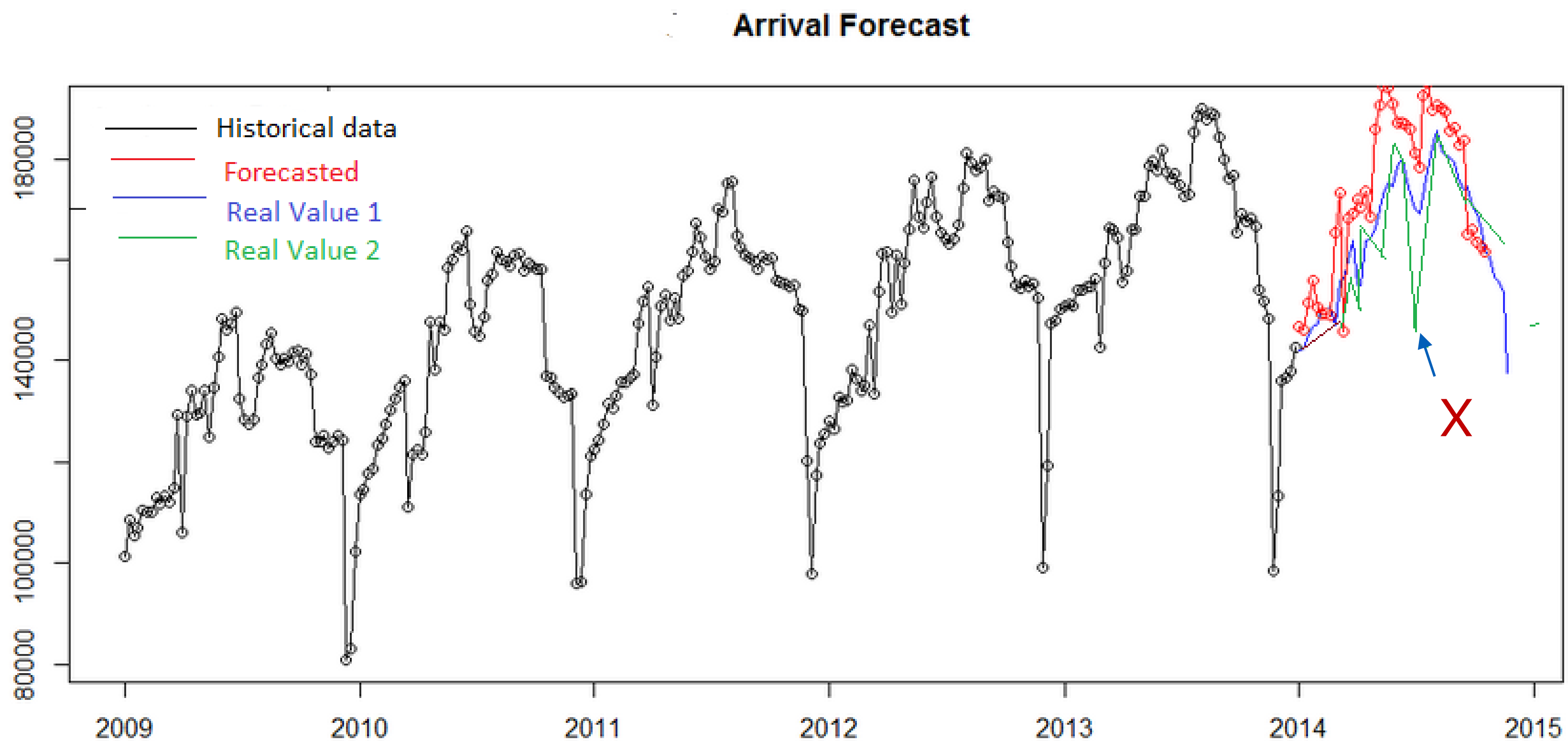
Outlier Analysis



_ X point is outlier, because:

- Its distance to the regression line is relatively far

Outlier Analysis



- For the time series above, the green time series contains an outlier because its distance to the forecasted one is relatively fare

Outlier Analysis

- See *Outlier Analysis* (Agarwal 2013), *An Introduction to Statistical Learning* (James et.al, 2014)

The diagram is for Illustration only, it is not derived from real data

3.2

Use Case 2: Customer Segmentation

Use Case 2: Customer Segmentation

_Customer

- Airlines, Marketing department. Not only E-Commerce

_Business Objective

- Optimize marketing campaign

_Business Requirements

- Segment the passengers based on their travel behavior

Use Case 2: Challenges

- The Challenges:
  - Want to segment all travelers, not only the ones who are already in airline loyalty program
  - Privacy: Anonymized result

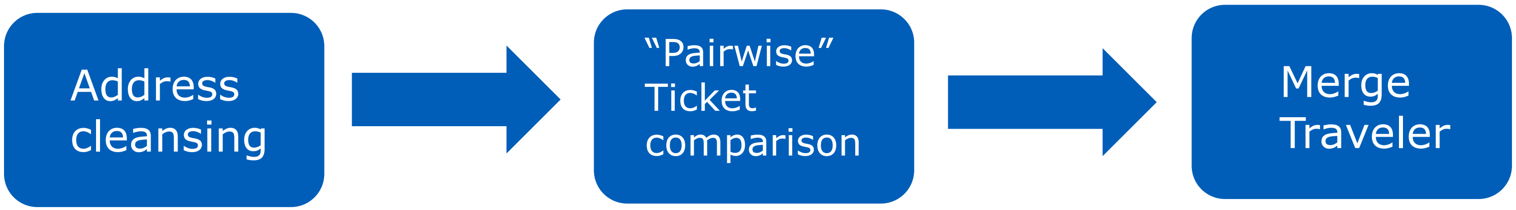
- Need to de-duplicate the same traveler based on their personal data: names, city of residence, phone number, zip code, email, gender, nationality, data of birth, id number, address, route, ...

Use Case 2: Data Nature

_The personal information is incomplete and very noisy:

- Names can be spell checked or reversed
- Addresses may change, or different
- City is not normalized: NY, NYC, New York, New York City
- Phone number is not normalized
- …

Use Case 2: Solution: Workflow



Use Case 2: Solution: Workflow

Address cleansing

Use gazetteer to normalize city, state, and country

Use Google Phone API to normalize phone

"Pairwise" Ticket comparison

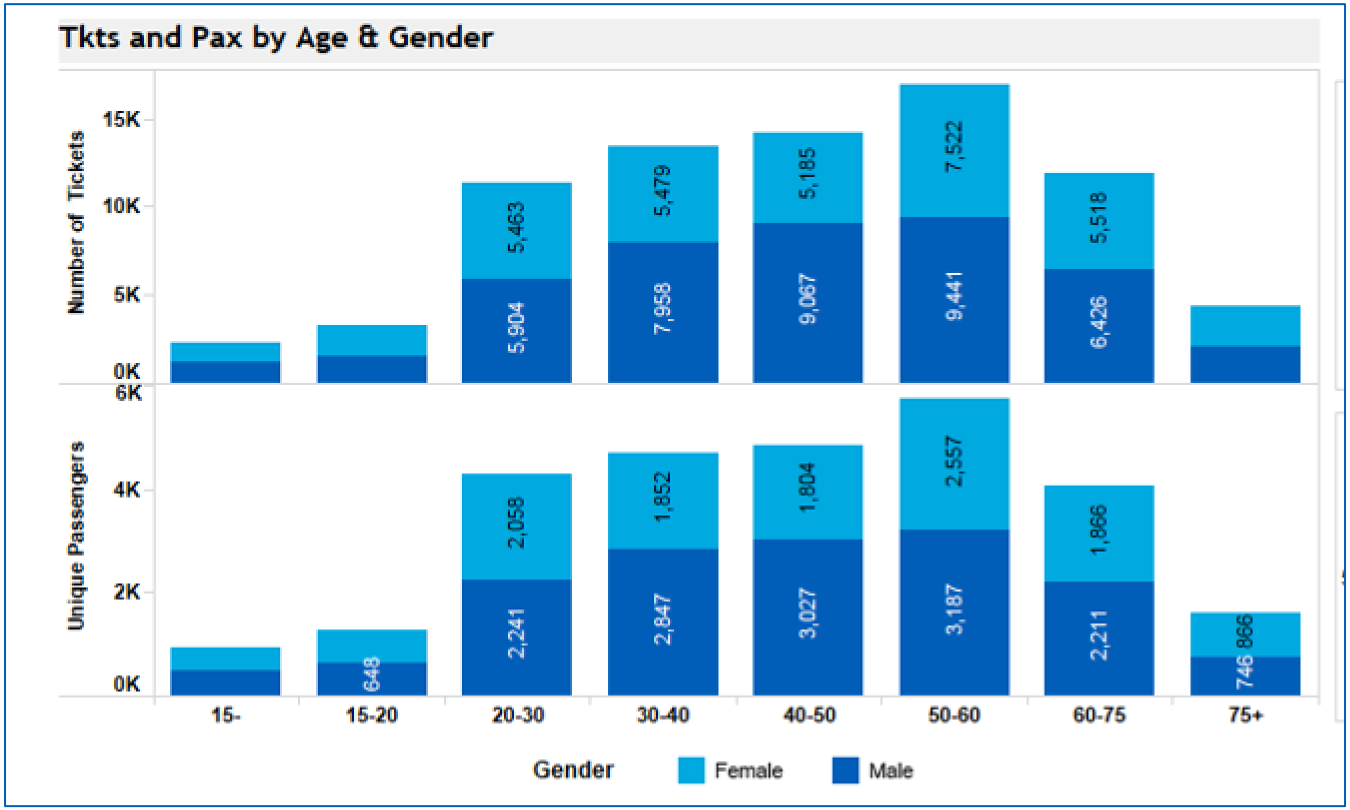
LSH and Min Hash followed by Connected Components of Graph

User Merge

Determine the final identity of the users: their residences, phone number, …

Segmenta tion

Use Case 2: Result



The diagram is for Illustration only, it is not derived from real data

Use Case 2: Segmentation

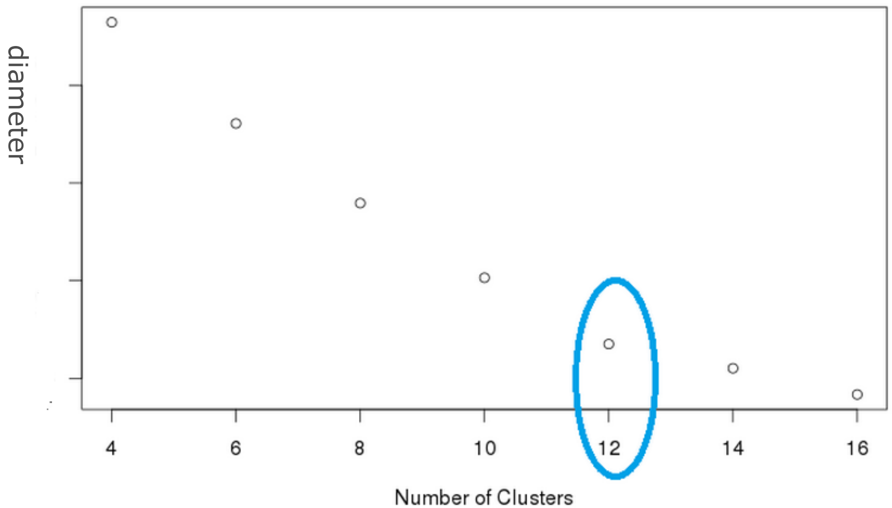
Features
Passenger Id
Number of travels the last 12 months
Average advance purchase
Average paid fare
Standard deviation of paid fare
Number of trips during working days
Ratio of repeated O&D
Domestic flight proportion
Age range
Gender
Nationality
Frequent flyer card level
Group bookings level
Family bookings level

The diagram is for Illustration only, it is not derived from real data

Use Case 2: Segmentation

_ Apply K-Means Clustering:

- Features Selection
  - e.g. Should we include gender / nationality ?
- K determination



The diagram is for Illustration only, it is not derived from real data

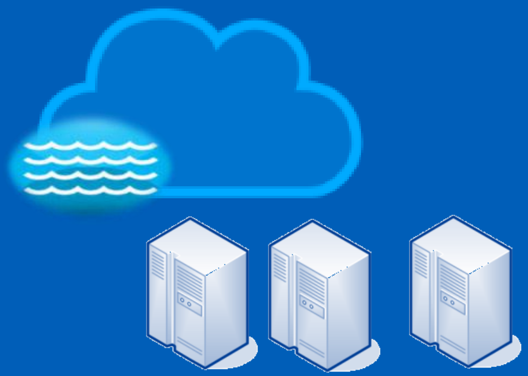
amadeus

Use Case 2: Segmentation

_ Interpret the result

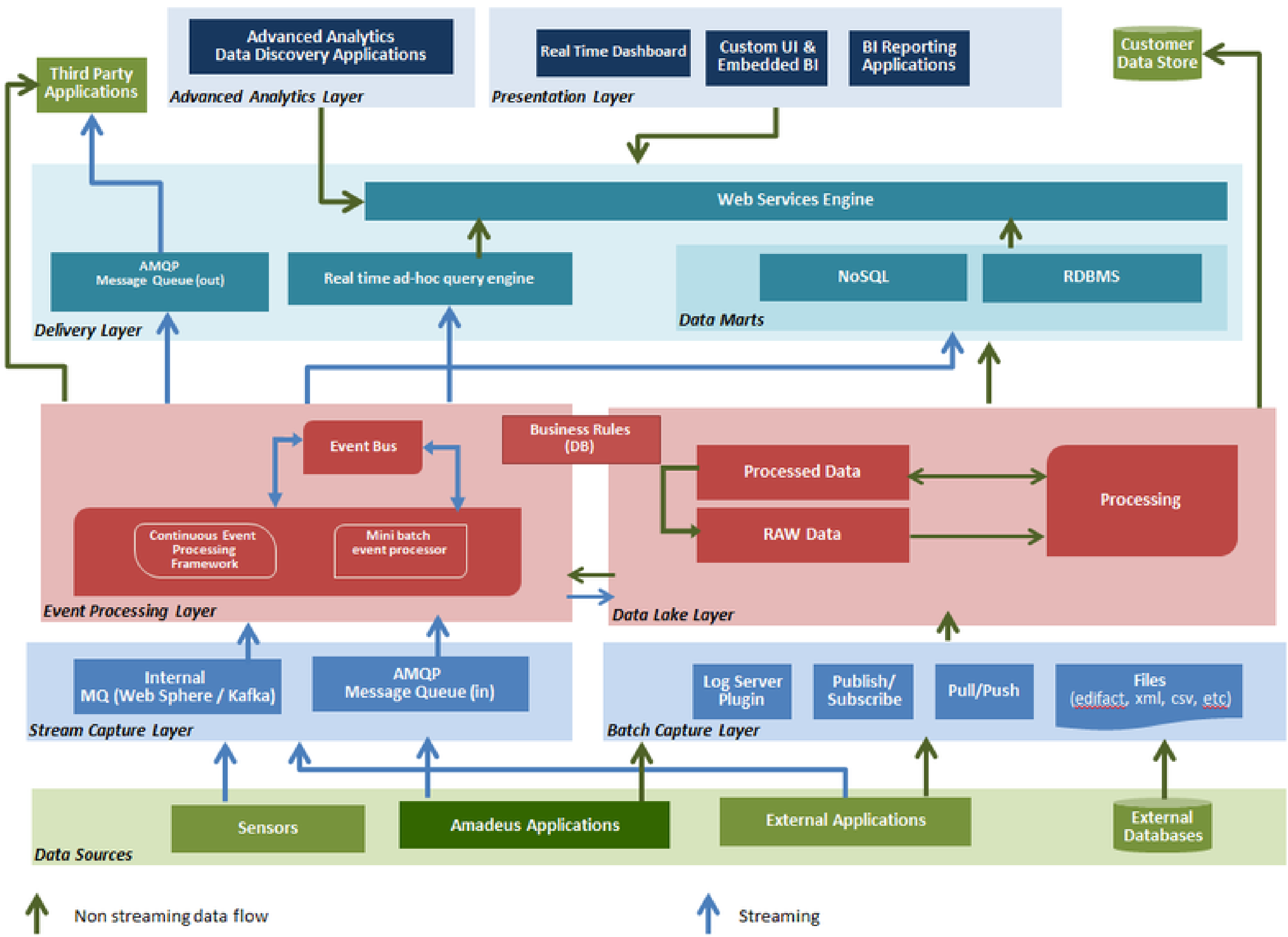
- Does the cluster repartition give some insights ?
  - e.g maybe it's not interesting to have clusters that cluster all men to one cluster and all women to another
- Is it reasonable to merge 'manually' some clusters to one ?

4 Technology Point of View



Amadeus Travel Intelligence Engine

Architecture Details



amadeus

Technology Used

Spark

Cloudera Impala

hadoop

SciPy

RStudio

Flink

Twitter algebird

Cassandra

Scoobi

5

Summary and Conclusion

- Does the cluster give some insights ?

Summary & Conclusion

_ Data Analysis plays important roles in travel industry

_ Analysis should start from what business actions need to be supported with data

_ We have seen two use cases:
- Conversion Rate Monitoring
- Customer Identification and then Segmentation
- Many Others

_ Data Preparation is costly and dominate the workflow of data analysis

_ Technology like Hadoop and Spark are helpful for doing data analysis at large scale

Thank you

You can follow us on:
AmadeusITGroup

amadeus.com/blog
amadeus.com