trivago

Large scale data processing pipelines at trivago: a use case

2016-11-15, Sevilla, Spain
Clemens Valiente

trivago

Clemens Valiente

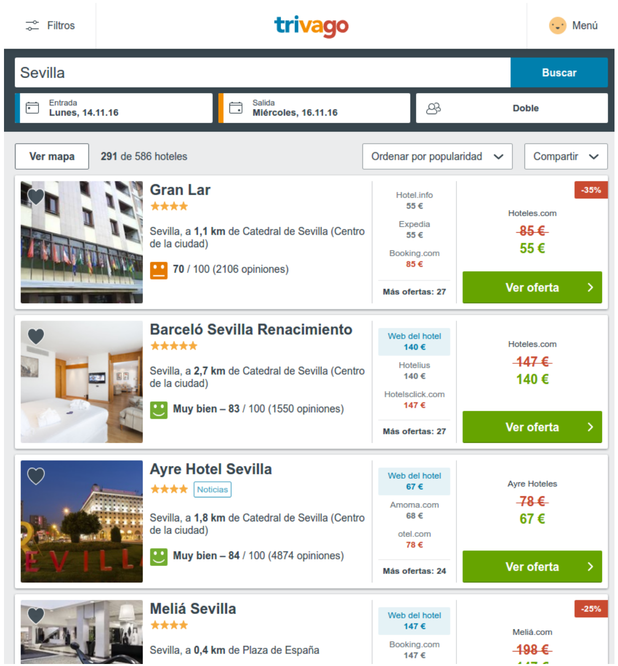
Senior Data Engineer
trivago Düsseldorf

Originally a mathematician
Studied at Uni Erlangen
At trivago for 5 years

Email: clemens.valiente@trivago.com
in de.linkedin.com/in/clemensvalien

Data driven PR and External Communication

Price information collected from the various booking websites and shown to our visitors also gives us a thorough overview over trends and development of hotel prices. This knowledge then is used by our Content Marketing & Communication Department (CMC) to write stories and articles.



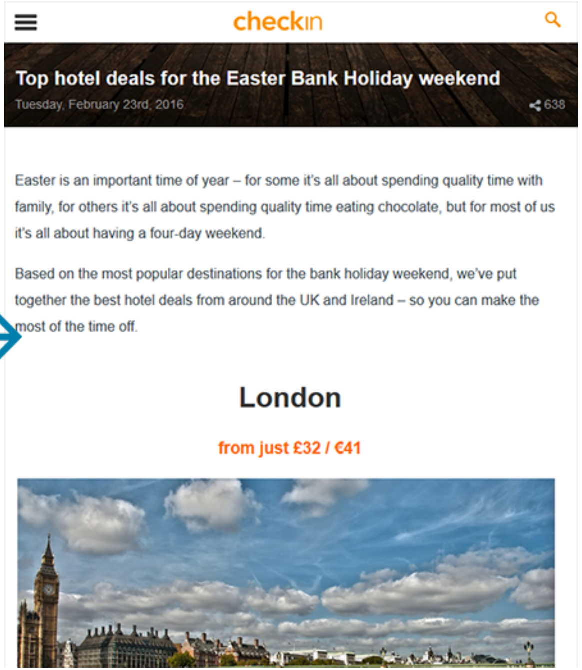
291 de 586 hoteles

Gran Lar ★★★★
Sevilla, a **1,1 km** de Catedral de Sevilla (Centro de la ciudad)
70 / 100 (2106 opiniones)
Hotel.info 55 € · Expedia 55 € · Booking.com 85 €
Más ofertas: 27
-35% Hoteles.com ~~85 €~~ 55 €

Barceló Sevilla Renacimiento ★★★★★
Sevilla, a **2,7 km** de Catedral de Sevilla (Centro de la ciudad)
Muy bien – 83 / 100 (1550 opiniones)
Web del hotel 140 € · Hotelius 140 € · Hotelsclick.com 147 €
Más ofertas: 27
Hoteles.com ~~147 €~~ 140 €

Ayre Hotel Sevilla ★★★★ Noticias
Sevilla, a **1,8 km** de Catedral de Sevilla (Centro de la ciudad)
Muy bien – 84 / 100 (4874 opiniones)
Web del hotel 67 € · Amoma.com 68 € · otel.com 78 €
Más ofertas: 24
Ayre Hoteles ~~78 €~~ 67 €

Meliá Sevilla ★★★★
Sevilla, a **0,4 km** de Plaza de España
Web del hotel 147 € · Booking.com 147 €
-25% Meliá.com ~~198 €~~



Top hotel deals for the Easter Bank Holiday weekend

Tuesday, February 23rd, 2016

Easter is an important time of year – for some it's all about spending quality time with family, for others it's all about spending quality time eating chocolate, but for most of us it's all about having a four-day weekend.

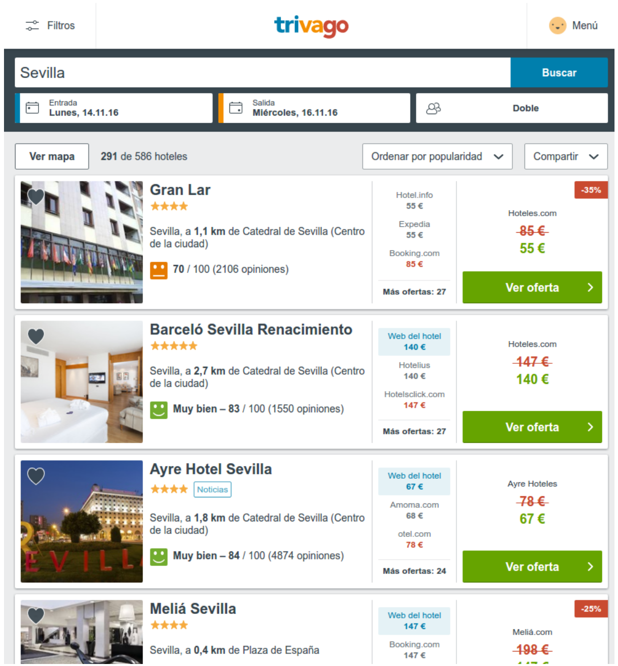
Based on the most popular destinations for the bank holiday weekend, we've put together the best hotel deals from around the UK and Ireland – so you can make the most of the time off.

London

from just £32 / €41

Data driven PR and External Communication

Price information collected from the various booking websites and shown to our visitors also gives us a thorough overview over trends and development of hotel prices. This knowledge then is used by our Content Marketing & Communication Department (CMC) to write stories and articles.



Sevilla

Entrada Lunes, 14.11.16 | Salida Miércoles, 16.11.16 | Doble

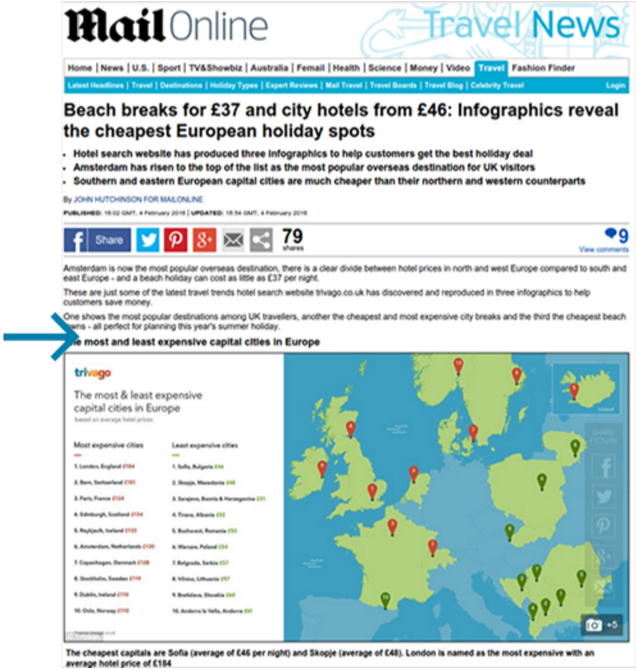
291 de 586 hoteles

Gran Lar ★★★★
Sevilla, a 1,1 km de Catedral de Sevilla (Centro de la ciudad)
70 / 100 (2106 opiniones)
Hotel.info 55 € | Expedia 55 € | Booking.com 85 €
Más ofertas: 27
-35% Hoteles.com ~~85 €~~ 55 €

Barceló Sevilla Renacimiento ★★★★★
Sevilla, a 2,7 km de Catedral de Sevilla (Centro de la ciudad)
Muy bien – 83 / 100 (1550 opiniones)
Web del hotel 140 € | Hotelius 140 € | Hotelsclick.com 147 €
Más ofertas: 27
Hoteles.com ~~147 €~~ 140 €

Ayre Hotel Sevilla ★★★★ Noticias
Sevilla, a 1,8 km de Catedral de Sevilla (Centro de la ciudad)
Muy bien – 84 / 100 (4874 opiniones)
Web del hotel 67 € | Amoma.com 68 € | otel.com 78 €
Más ofertas: 24
Ayre Hoteles ~~78 €~~ 67 €

Meliá Sevilla ★★★★
Sevilla, a 0,4 km de Plaza de España
Web del hotel 147 € | Booking.com 147 €
-25% Meliá.com ~~198 €~~


MailOnline
Travel News
Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money | Video | Travel | Fashion Finder
Latest Headlines | Travel | Destinations | Holiday Types | Expert Reviews | Mail Travel | Travel Boards | Travel Blog | Celebrity Travel
Login
Share
79 shares
9
View comments
trivago
The most & least expensive capital cities in Europe
based on average hotel prices
Most expensive cities
1. London, England £184
2. Bern, Switzerland £151
3. Paris, France £134
4. Edinburgh, Scotland £134
5. Reykjavik, Iceland £133
6. Amsterdam, Netherlands £130
7. Copenhagen, Denmark £128
8. Stockholm, Sweden £119
9. Dublin, Ireland £110
10. Oslo, Norway £110
Least expensive cities
1. Sofia, Bulgaria £46
2. Skopje, Macedonia £48
3. Sarajevo, Bosnia & Herzegovina £51
4. Tirana, Albania £52
5. Bucharest, Romania £53
6. Warsaw, Poland £54
7. Belgrade, Serbia £57
8. Vilnius, Lithuania £57
9. Bratislava, Slovakia £60
10. Andorra la Vella, Andorra £61
+5

Beach breaks for £37 and city hotels from £46: Infographics reveal the cheapest European holiday spots

- **Hotel search website has produced three infographics to help customers get the best holiday deal**
- **Amsterdam has risen to the top of the list as the most popular overseas destination for UK visitors**
- **Southern and eastern European capital cities are much cheaper than their northern and western counterparts**

By JOHN HUTCHINSON FOR MAILONLINE

PUBLISHED: 18:02 GMT, 4 February 2016 | **UPDATED:** 18:54 GMT, 4 February 2016

Amsterdam is now the most popular overseas destination, there is a clear divide between hotel prices in north and west Europe compared to south and east Europe - and a beach holiday can cost as little as £37 per night.

These are just some of the latest travel trends hotel search website trivago.co.uk has discovered and reproduced in three infographics to help customers save money.

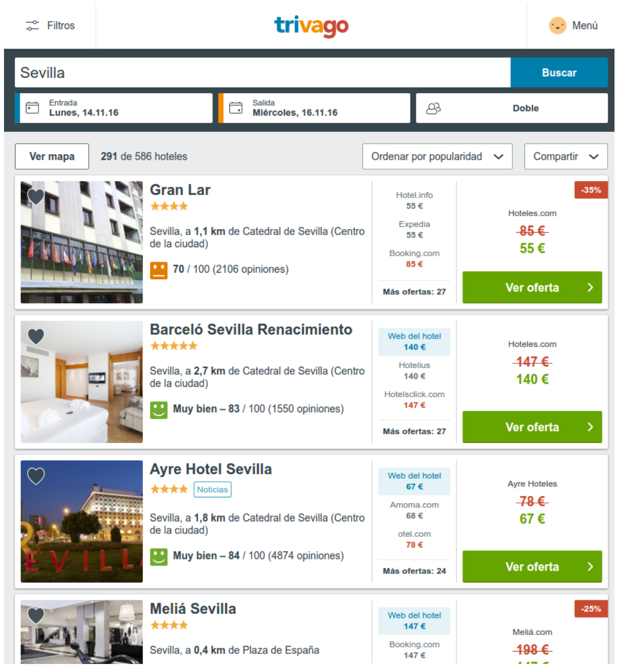
One shows the most popular destinations among UK travellers, another the cheapest and most expensive city breaks and the third the cheapest beach towns - all perfect for planning this year's summer holiday.

The most and least expensive capital cities in Europe

The cheapest capitals are Sofia (average of £46 per night) and Skopje (average of £48). London is named as the most expensive with an average hotel price of £184

Data driven PR and External Communication

Price information collected from the various booking websites and shown to our visitors also gives us a thorough overview over trends and development of hotel prices. This knowledge then is used by our Content Marketing & Communication Department (CMC) to write stories and articles.



Sevilla

Entrada Lunes, 14.11.16

Salida Miércoles, 16.11.16

Doble

291 de 586 hoteles

Gran Lar ★★★★

Sevilla, a **1,1 km** de Catedral de Sevilla (Centro de la ciudad)

70 / 100 (2106 opiniones)

Hotel.info 55 €
Expedia 55 €
Booking.com 85 €
Más ofertas: 27

-35% Hoteles.com ~~85 €~~ 55 €

Barceló Sevilla Renacimiento ★★★★★

Sevilla, a **2,7 km** de Catedral de Sevilla (Centro de la ciudad)

Muy bien – 83 / 100 (1550 opiniones)

Web del hotel 140 €
Hotelius 140 €
Hotelsclick.com 147 €
Más ofertas: 27

Hoteles.com ~~147 €~~ 140 €

Ayre Hotel Sevilla ★★★★

Sevilla, a **1,8 km** de Catedral de Sevilla (Centro de la ciudad)

Muy bien – 84 / 100 (4874 opiniones)

Web del hotel 67 €
Amoma.com 68 €
otel.com 78 €
Más ofertas: 24

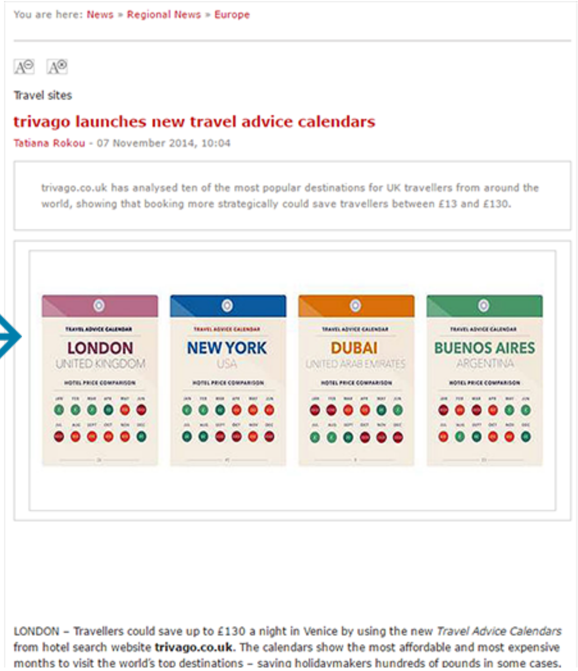
Ayre Hoteles ~~78 €~~ 67 €

Meliá Sevilla ★★★★

Sevilla, a **0,4 km** de Plaza de España

Web del hotel 147 €
Booking.com 147 €

-25% Meliá.com ~~198 €~~



You are here: News » Regional News » Europe

Travel sites

trivago launches new travel advice calendars

Tatiana Rokou - 07 November 2014, 10:04

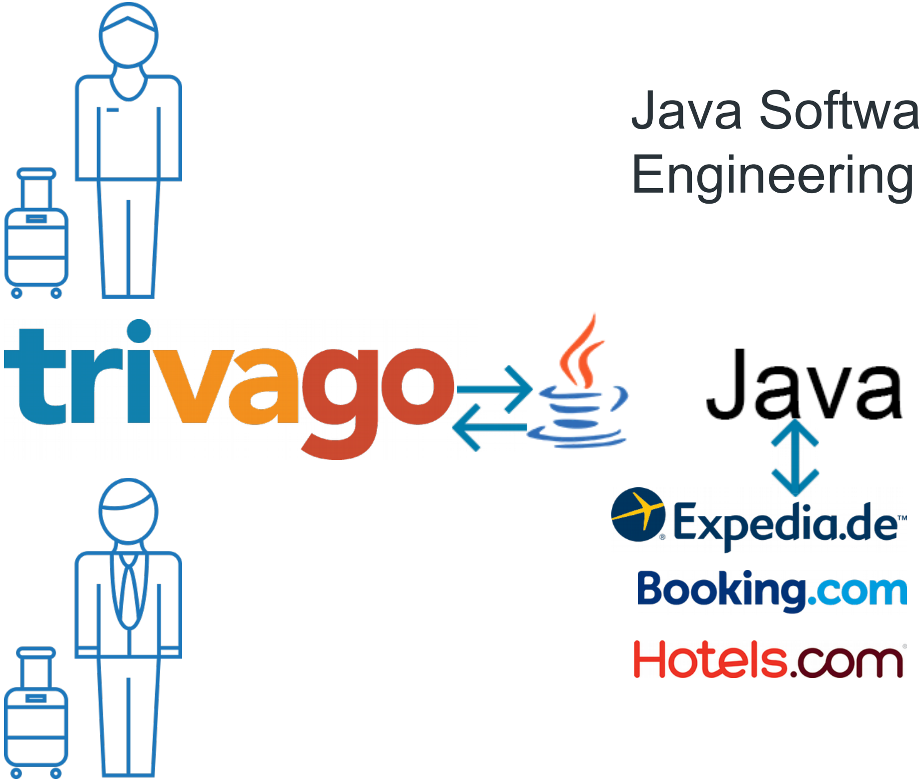
trivago.co.uk has analysed ten of the most popular destinations for UK travellers from around the world, showing that booking more strategically could save travellers between £13 and £130.

LONDON – Travellers could save up to £130 a night in Venice by using the new *Travel Advice Calendars* from hotel search website **trivago.co.uk**. The calendars show the most affordable and most expensive months to visit the world's top destinations – saving holidaymakers hundreds of pounds in some cases.

The past: Data pipeline 2010 – 2015

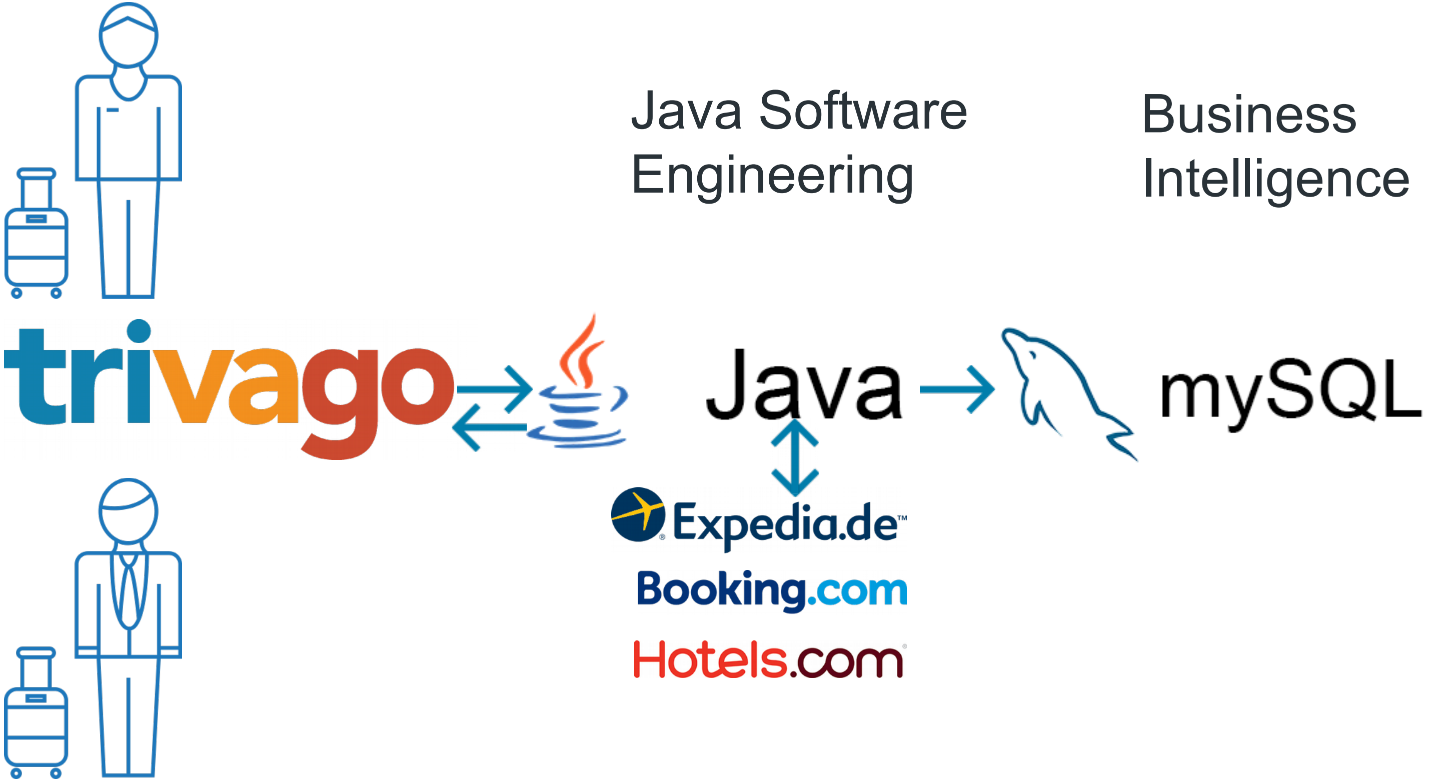
trivago

The past: Data pipeline 2010 – 2015



Java Software Engineering

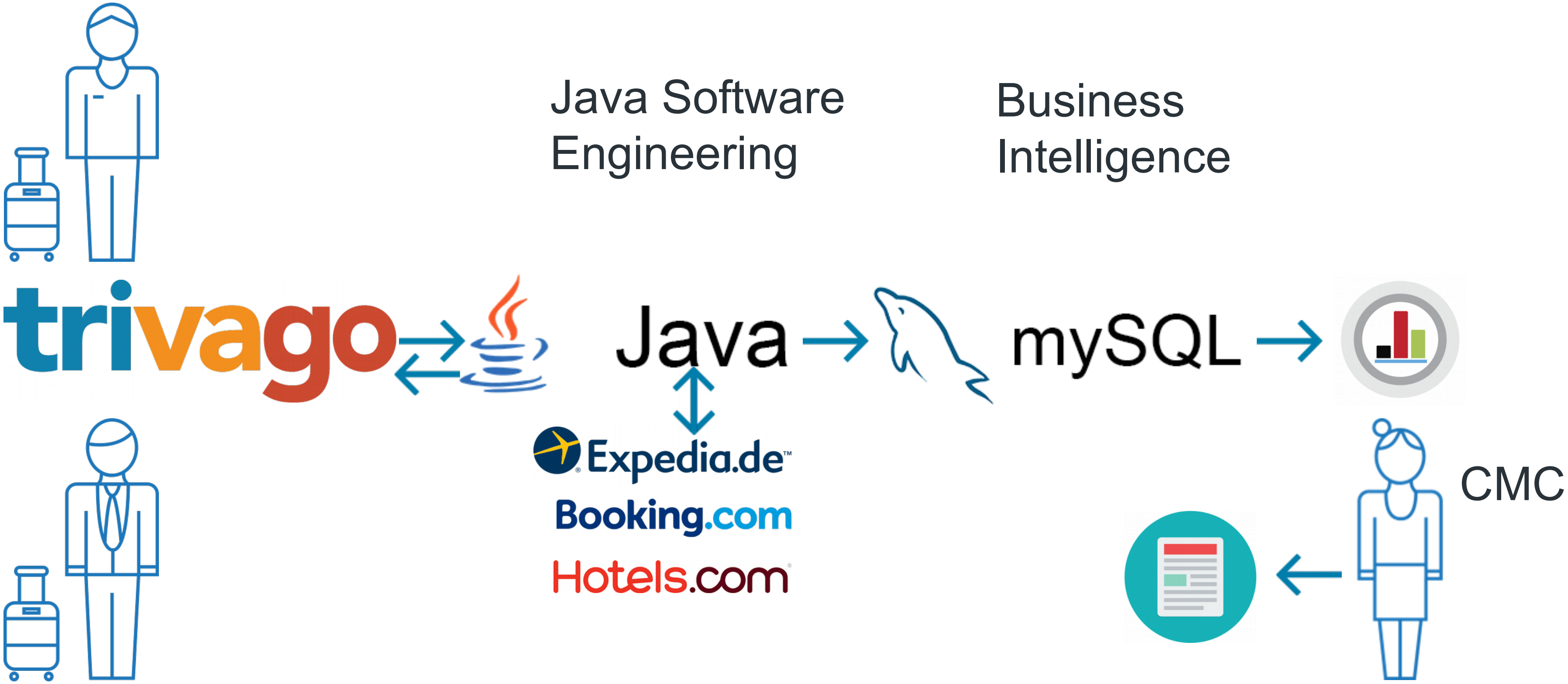
The past: Data pipeline 2010 – 2015



Java Software Engineering

Business Intelligence

The past: Data pipeline 2010 – 2015



Java Software Engineering

Business Intelligence

CMC

The past: Data pipeline 2010 – 2015
Facts & Figures

Price dimensions

- Around one million hotels
- 250 booking websites
- Travellers search for up to 180 days in advance
- Data collected over five years

The past: Data pipeline 2010 – 2015 Facts & Figures

Price dimensions

- Around one million hotels
- 250 booking websites
- Travellers search for up to 180 days in advance
- Data collected over five years

Restrictions

- Only single night stays
- Only prices from European visitors
- Prices cached up to 30 minutes
- One price per hotel, website and arrival date per day
- “Insert ignore”: The first price per key wins

The past: Data pipeline 2010 – 2015
Facts & Figures

Price dimensions

- Around one million hotels
- 250 booking websites
- Travellers search for up to 180 days in advance
- Data collected over five years

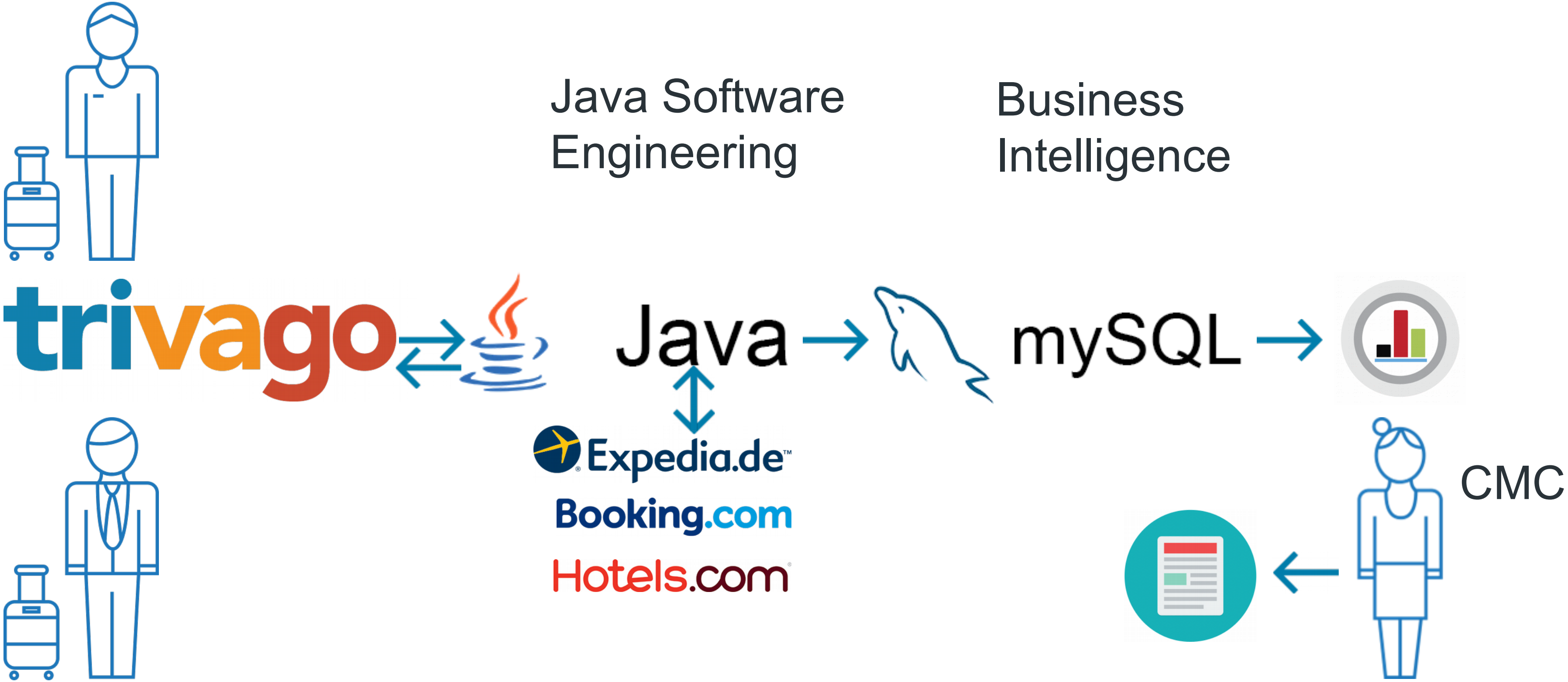
Restrictions

- Only single night stays
- Only prices from European visitors
- Prices cached up to 30 minutes
- One price per hotel, website and arrival date per day
- “Insert ignore”: The first price per key wins

Size of data

- We collected a total of 56 billion prices in those five years
- Towards the end of this pipeline in early 2015 on average around 100 million prices per day were written to BI

The past: Data pipeline 2010 – 2015

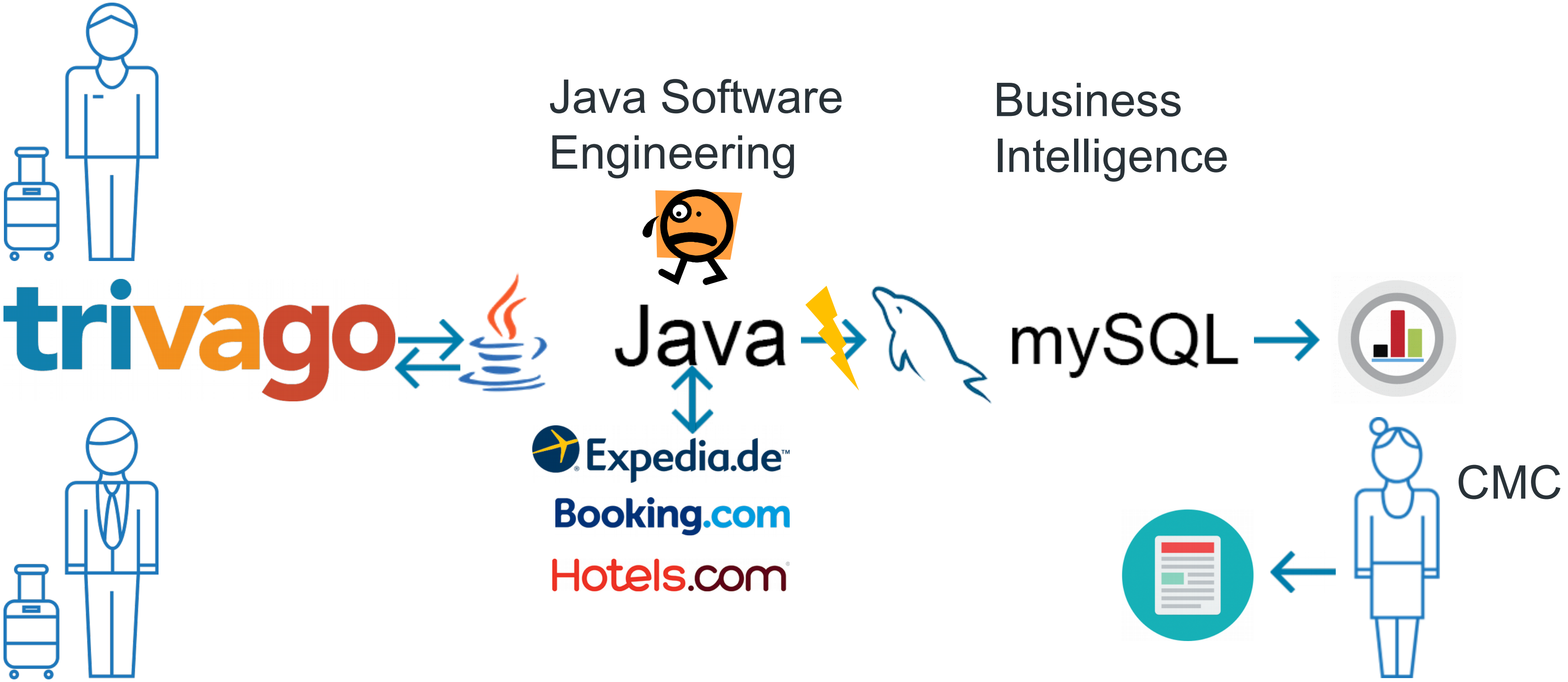


Java Software Engineering

Business Intelligence

CMC

The past: Data pipeline 2010 – 2015

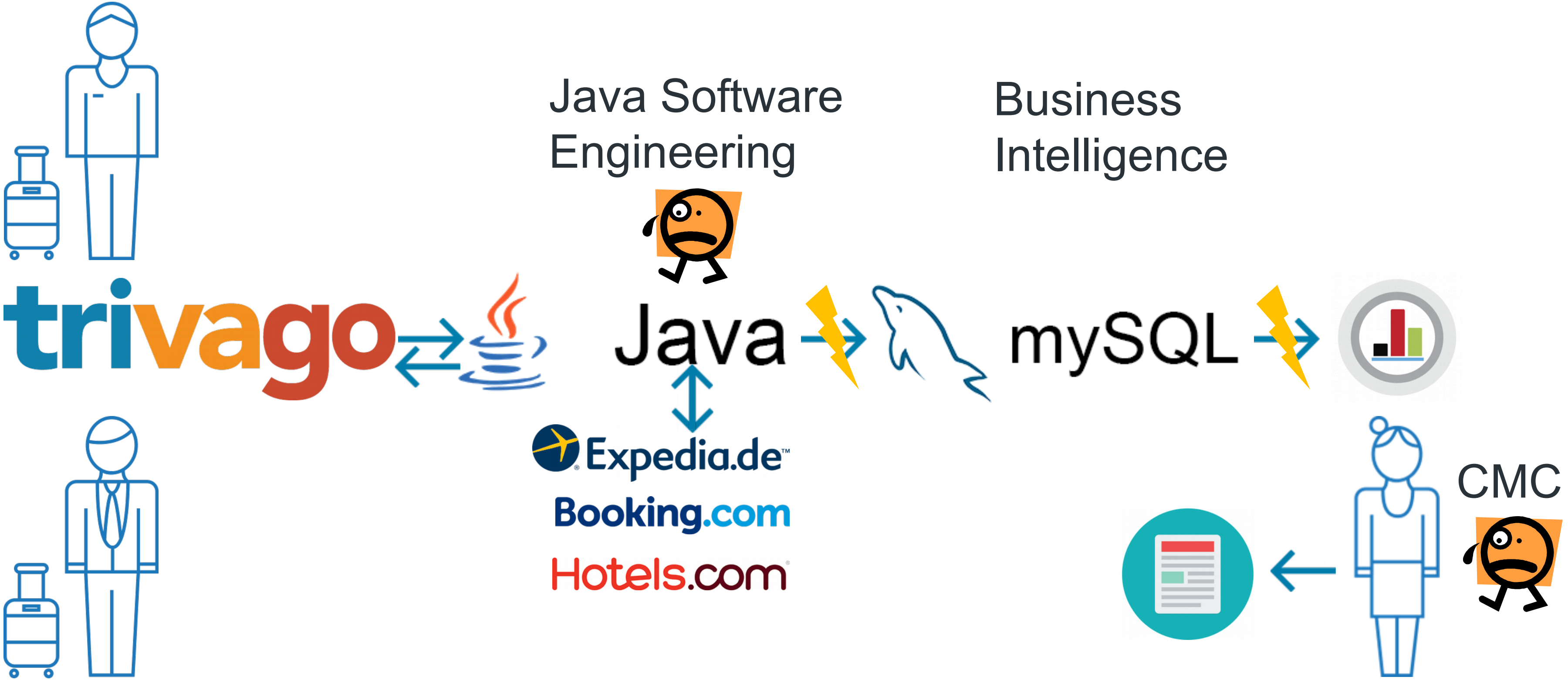


Java Software Engineering

Business Intelligence

CMC

The past: Data pipeline 2010 – 2015

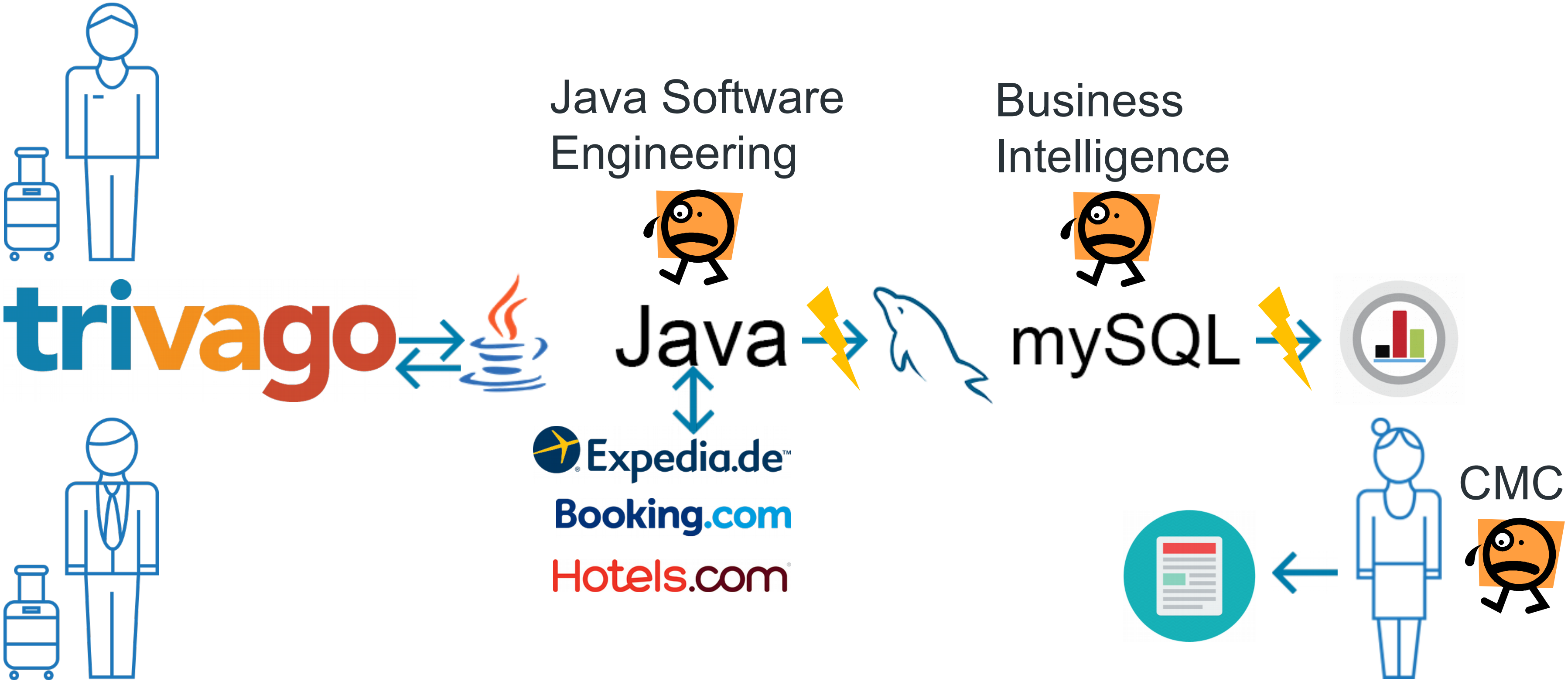


Java Software Engineering

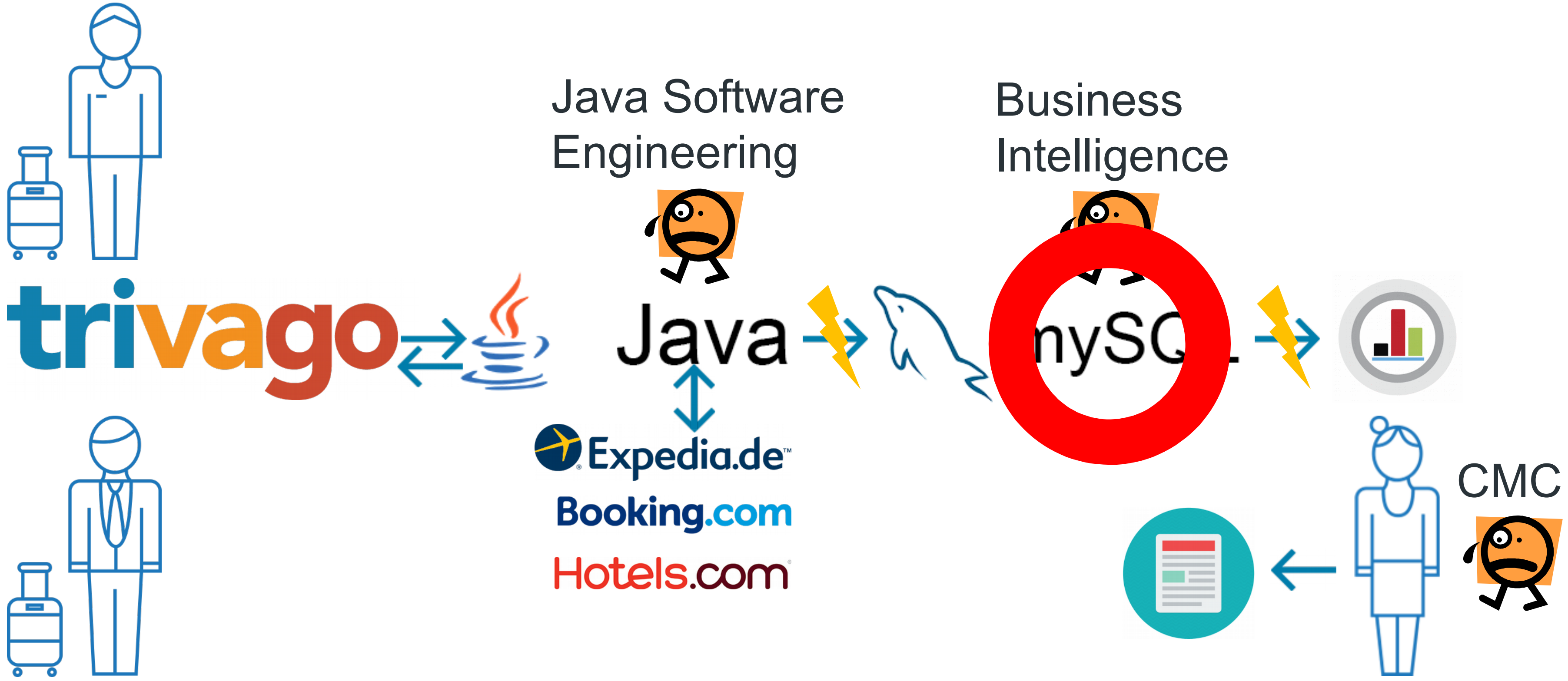
Business Intelligence

CMC

The past: Data pipeline 2010 – 2015



The past: Data pipeline 2010 – 2015



Java Software Engineering

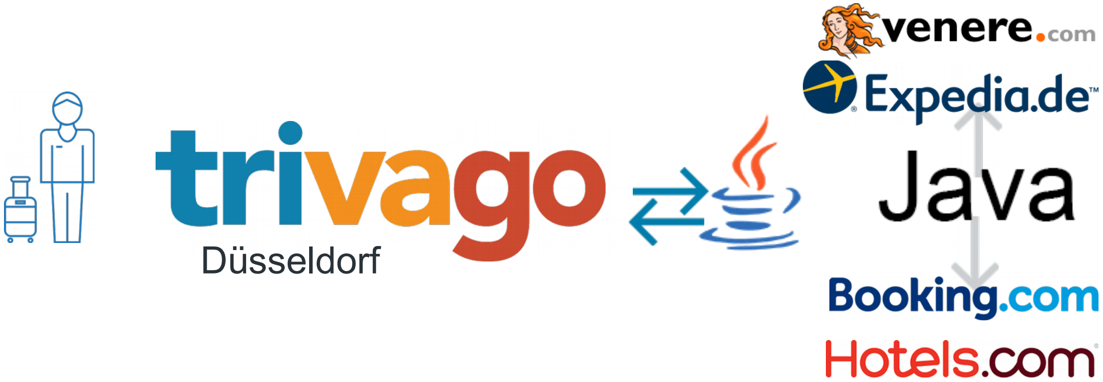
Business Intelligence

CMC

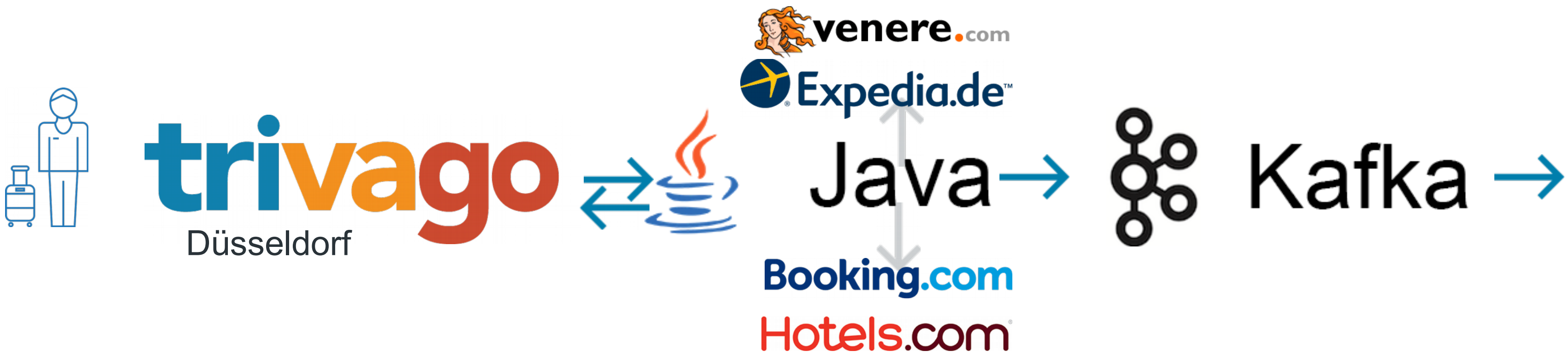
Refactoring the pipeline: Requirements

- Scales with an arbitrary amount of data (future proof)
- reliable and resilient
- low performance impact on Java backend
- long term storage of raw input data
- fast processing of filtered and aggregated data
- Open source
- we want to log everything:
    - more prices
        - Length of stay, room type, breakfast info, room category, domain
    - with more information
        - Net & gross price, city tax, resort fee, affiliate fee, VAT

Present data pipeline 2016 – ingestion



Present data pipeline 2016 – ingestion

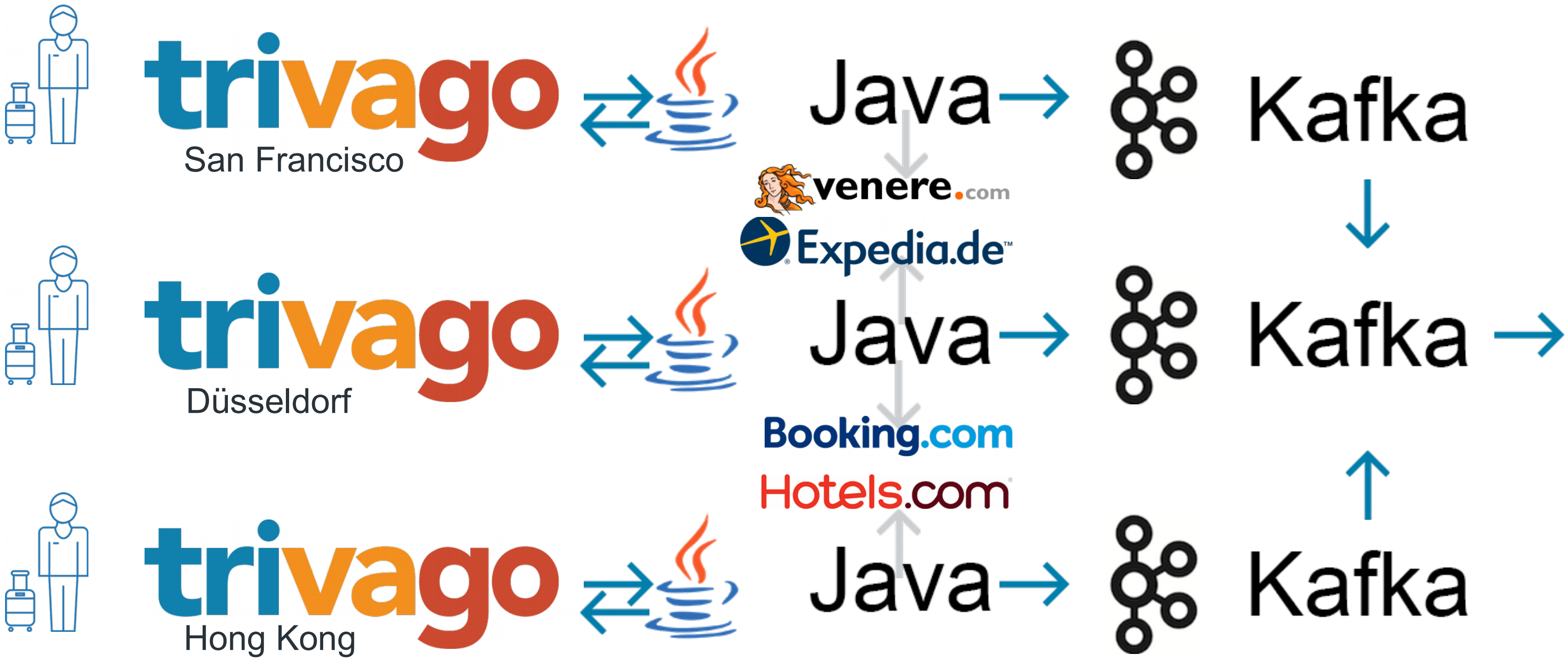


Düsseldorf

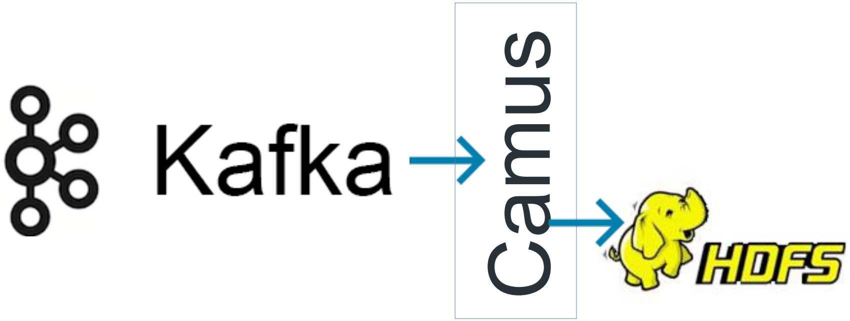
Java

Kafka

Present data pipeline 2016 – ingestion



Present data pipeline 2016 – processing

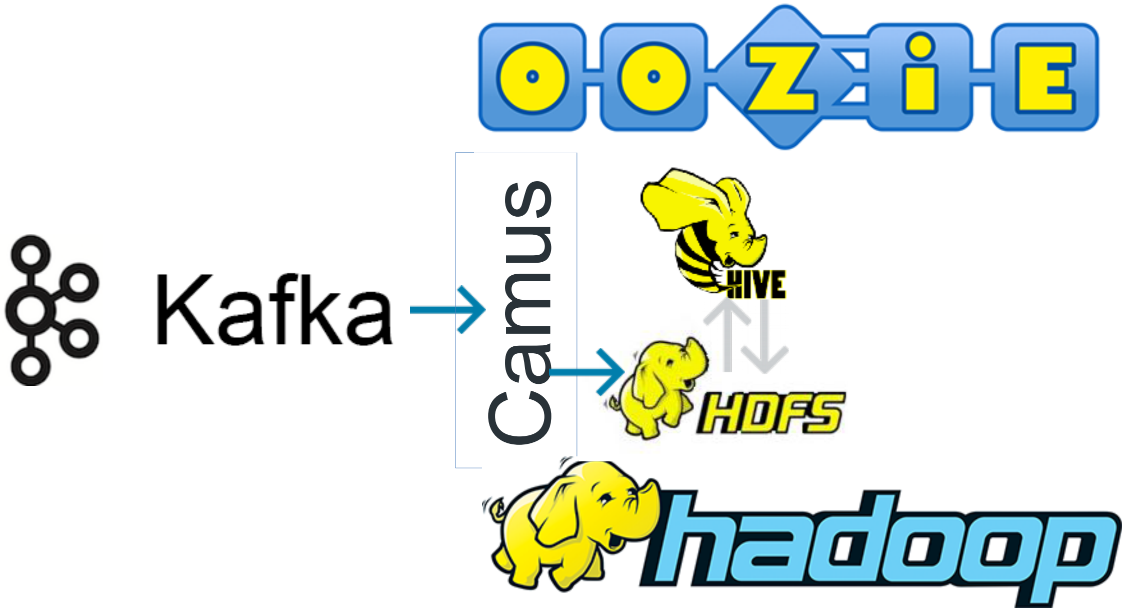


Kafka

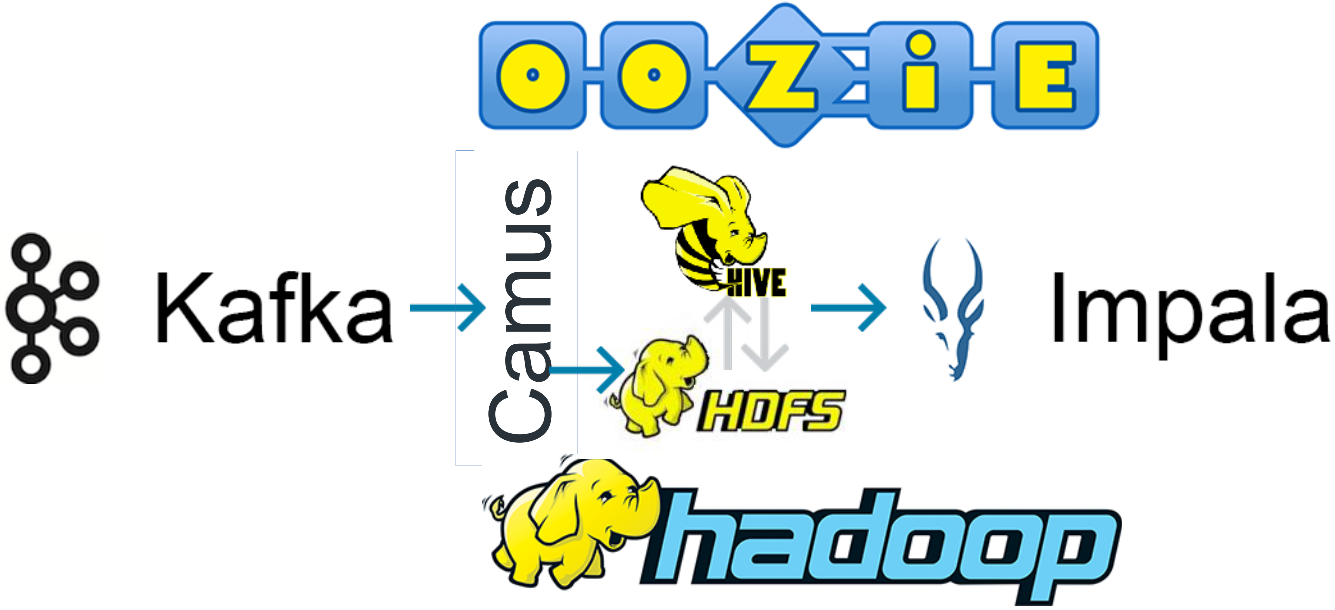
Camus

HDFS

Present data pipeline 2016 – processing



Present data pipeline 2016 – processing

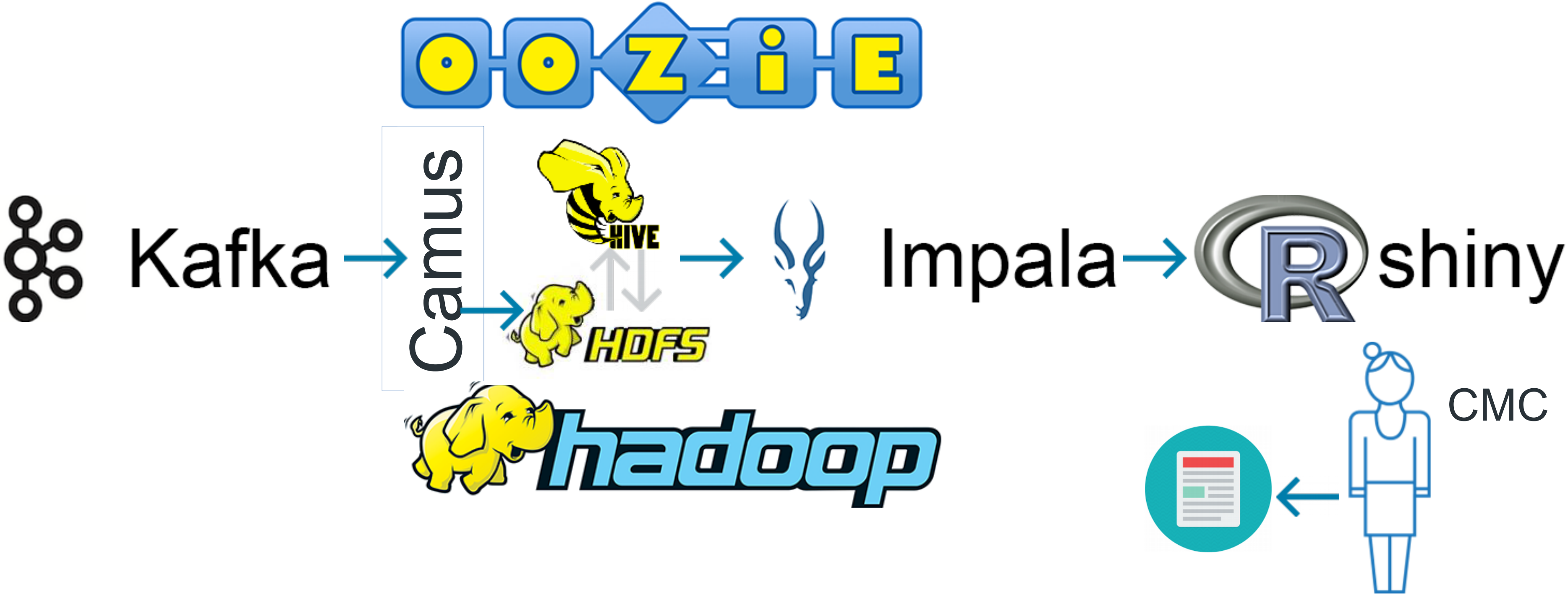


Kafka

Camus

Impala

Present data pipeline 2016 – processing



Kafka

Camus

Impala

shiny

CMC

Present data pipeline 2016 – facts & figures

Cluster specifications

- 51 machines
- 1.7 PB disc space, 60% used
- 3.6 TB memory in Yarn
- 1440 VCores (24-32 Cores per machine)

Present data pipeline 2016 – facts & figures

Cluster specifications

- 51 machines
- 1.7 PB disc space, 60% used
- 3.6 TB memory in Yarn
- 1440 VCores (24-32 Cores per machine)

Data Size (price log)

- 2.6 trillion messages collected so far
- 7 billion messages/day
- 160 TB of data

Present data pipeline 2016 – facts & figures

Cluster specifications

- 51 machines
- 1.7 PB disc space, 60% used
- 3.6 TB memory in Yarn
- 1440 VCores (24-32 Cores per machine)

Data Size (price log)

- 2.6 trillion messages collected so far
- 7 billion messages/day
- 160 TB of data

Data processing

- Camus: 30 mappers writing data in 10 minute intervals
- First aggregation/filtering stage in Hive runs in 30 minutes with 5 days of CPU time spent
- Impala Queries across >100 GB of result tables usually done within a few seconds

Present data pipeline 2016 – results after one and a half years in production

- Very reliable, barely any downtime or service interuptions of the system
- Java team is very happy – less load on their system
- BI team is very happy – more data, more ressources to process it
- CMC team is very happy
    - Faster results
    - Better quality of results due to more data
    - More detailed results
    - => Shorter research phase, more and better stories
    - => Less requests & workload for BI

Present data pipeline 2016 – use cases & status quo

Uses for price information

- Monitoring price parity in hotel market
- Anomaly and fraud detection
- Price feed for online marketing
- Display of price development and delivering price alerts to website visitors

Present data pipeline 2016 – use cases & status quo

Uses for price information

- Monitoring price parity in hotel market
- Anomaly and fraud detection
- Price feed for online marketing
- Display of price development and delivering price alerts to website visitors

Other data sources and usage

- Clicklog information from our website and mobile app
- Used for marketing performance analysis, product tests, invoice generation etc

Present data pipeline 2016 – use cases & status quo

Uses for price information

- Monitoring price parity in hotel market
- Anomaly and fraud detection
- Price feed for online marketing
- Display of price development and delivering price alerts to website visitors

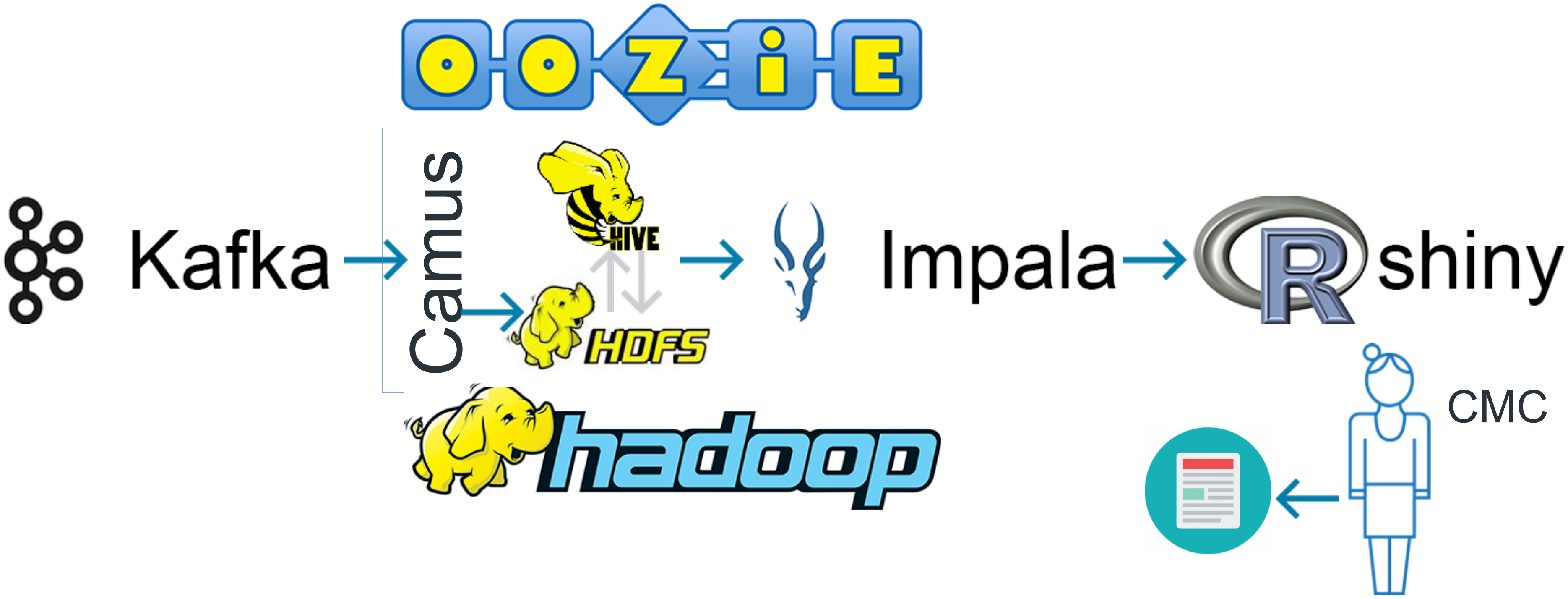
Other data sources and usage

- Clicklog information from our website and mobile app
- Used for marketing performance analysis, product tests, invoice generation etc

Status quo

- Our entire BI business logic runs on and through the kafka – hadoop pipeline
- Almost all departments rely on data, insights and metrics delivered by hadoop
- Most of the company could not do their job without hadoop data

Future data pipeline 2016/2017



Kafka

Camus

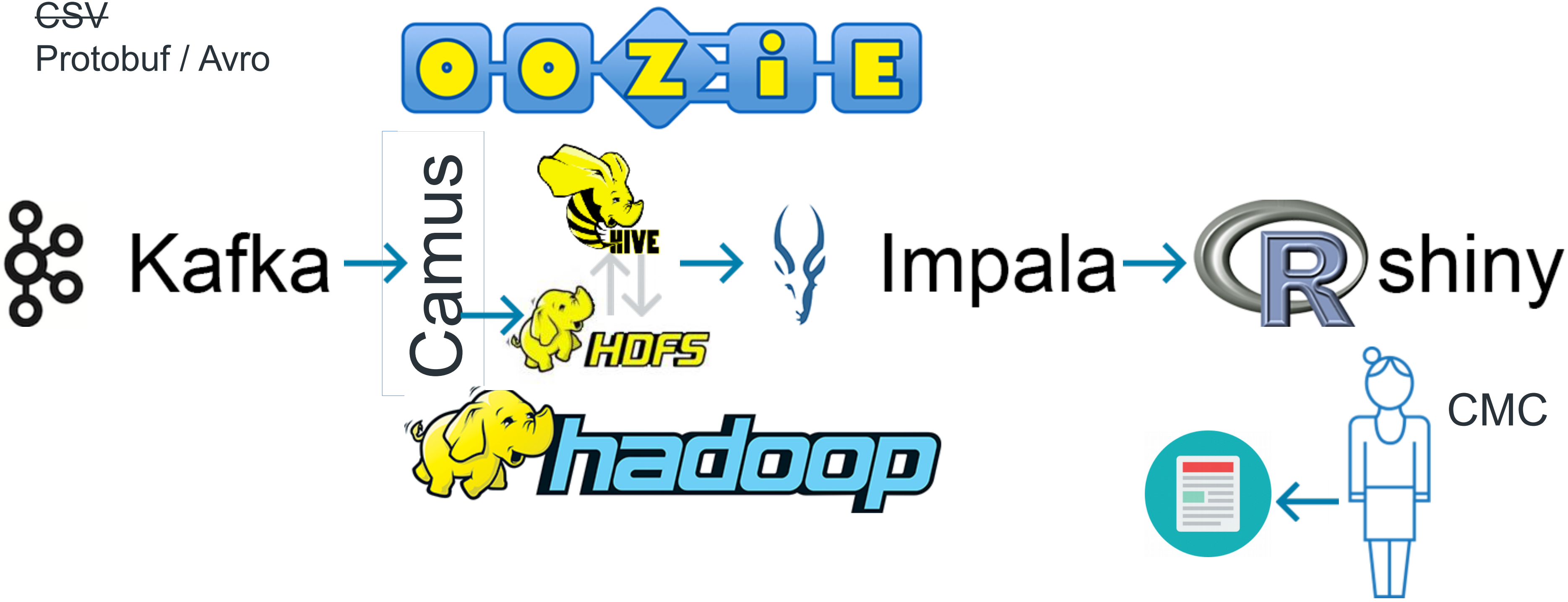
Impala

shiny

CMC

Future data pipeline 2016/2017

Message format:
~~CSV~~
Protobuf / Avro

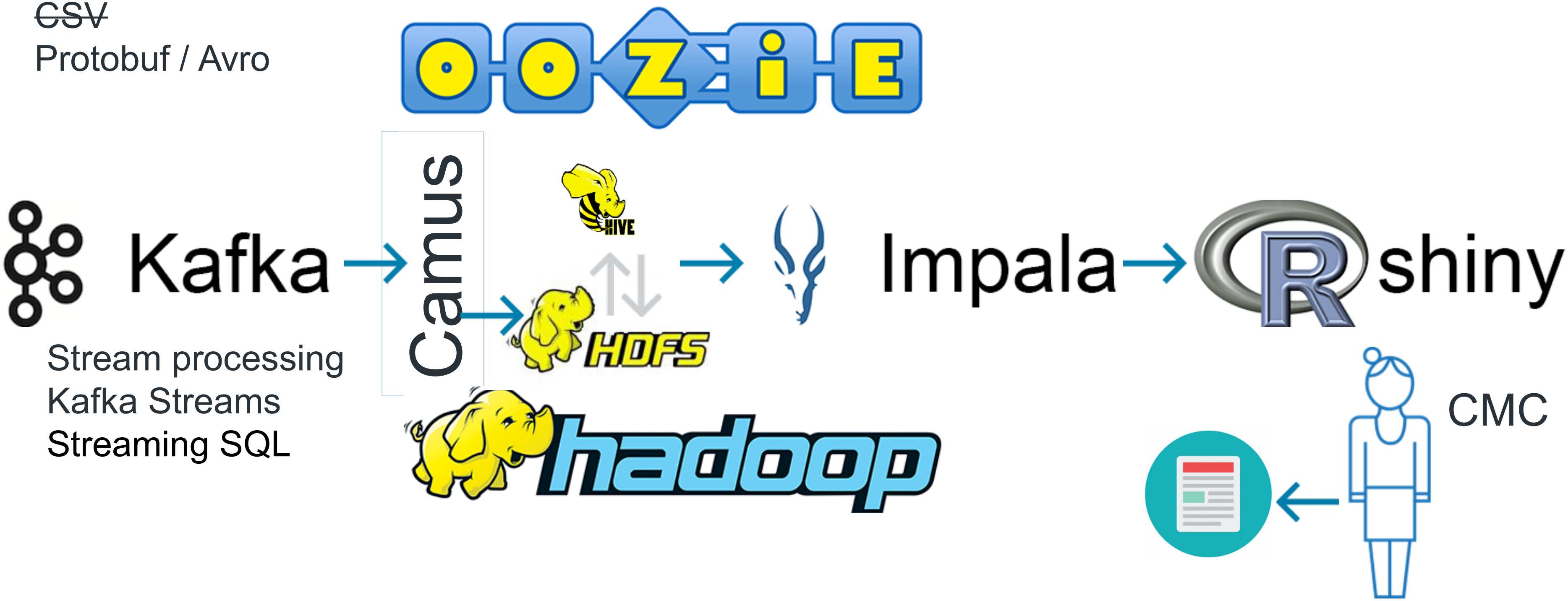


Kafka → Camus → Impala → shiny

CMC

Future data pipeline 2016/2017

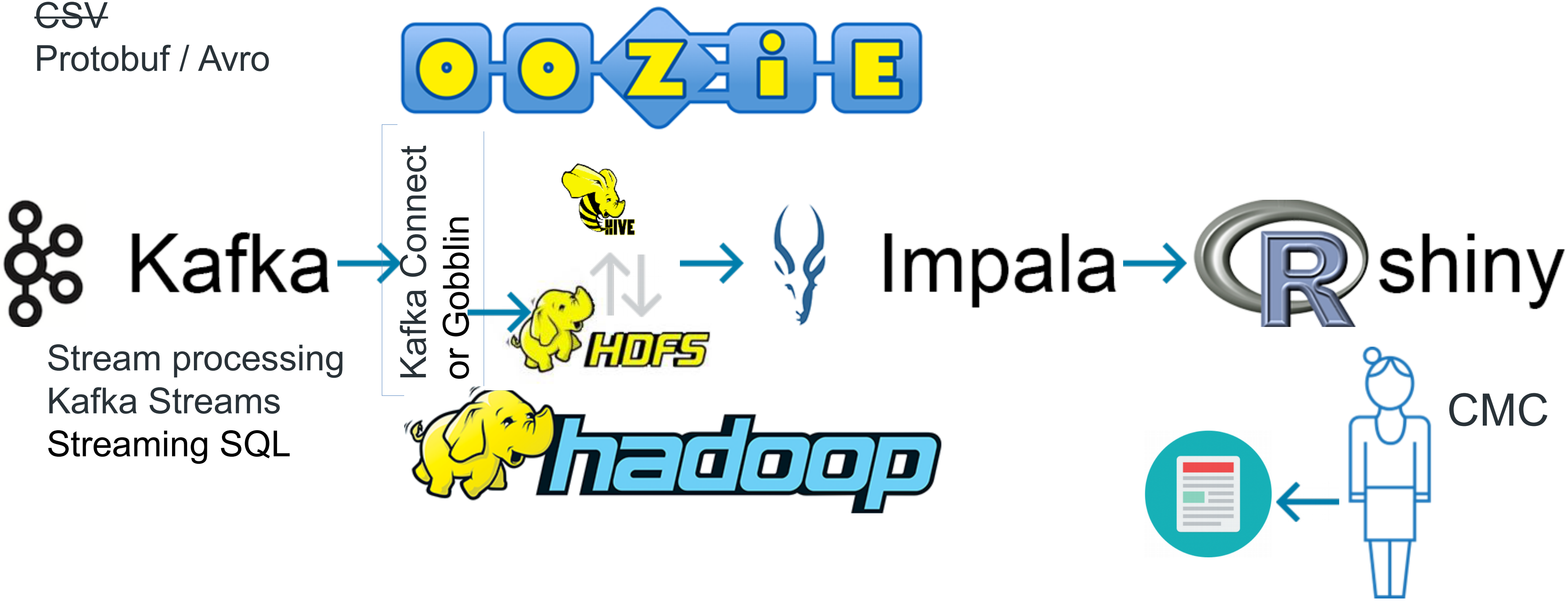
Message format:
~~CSV~~
Protobuf / Avro



Stream processing
Kafka Streams
Streaming SQL

Future data pipeline 2016/2017

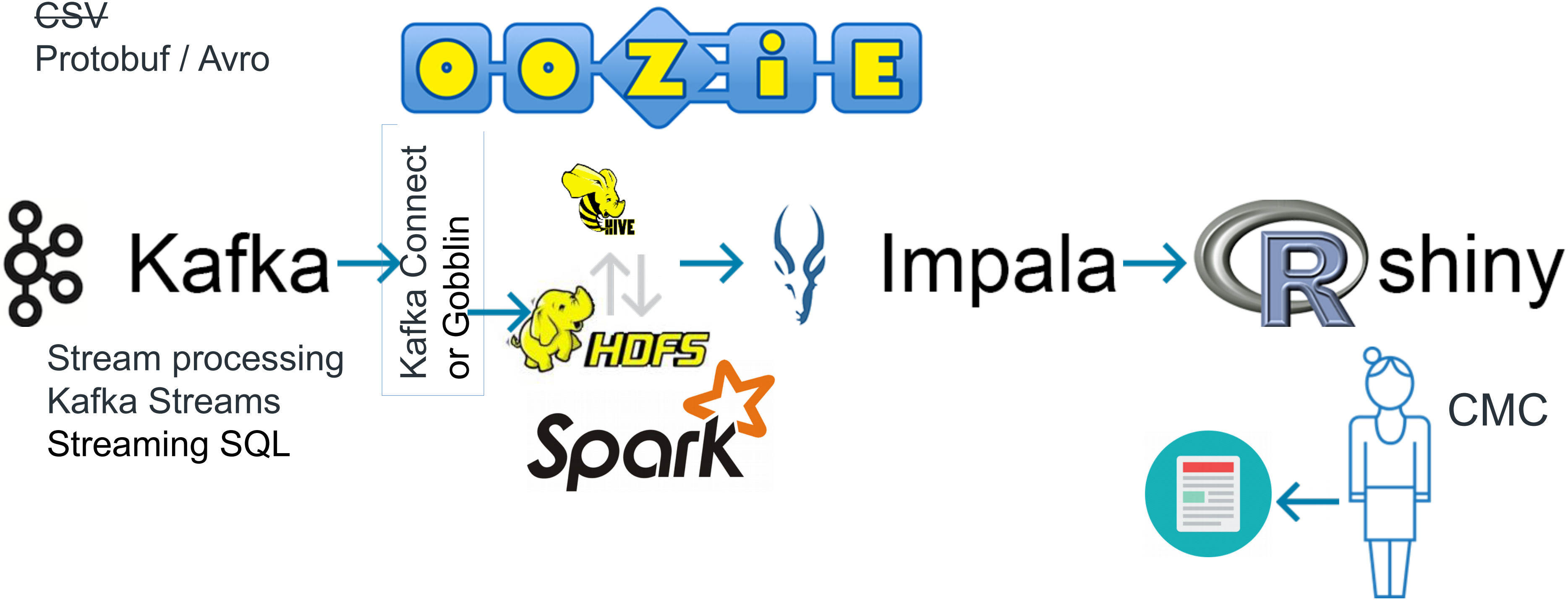
Message format:
~~CSV~~
Protobuf / Avro



Stream processing
Kafka Streams
Streaming SQL

Future data pipeline 2016/2017

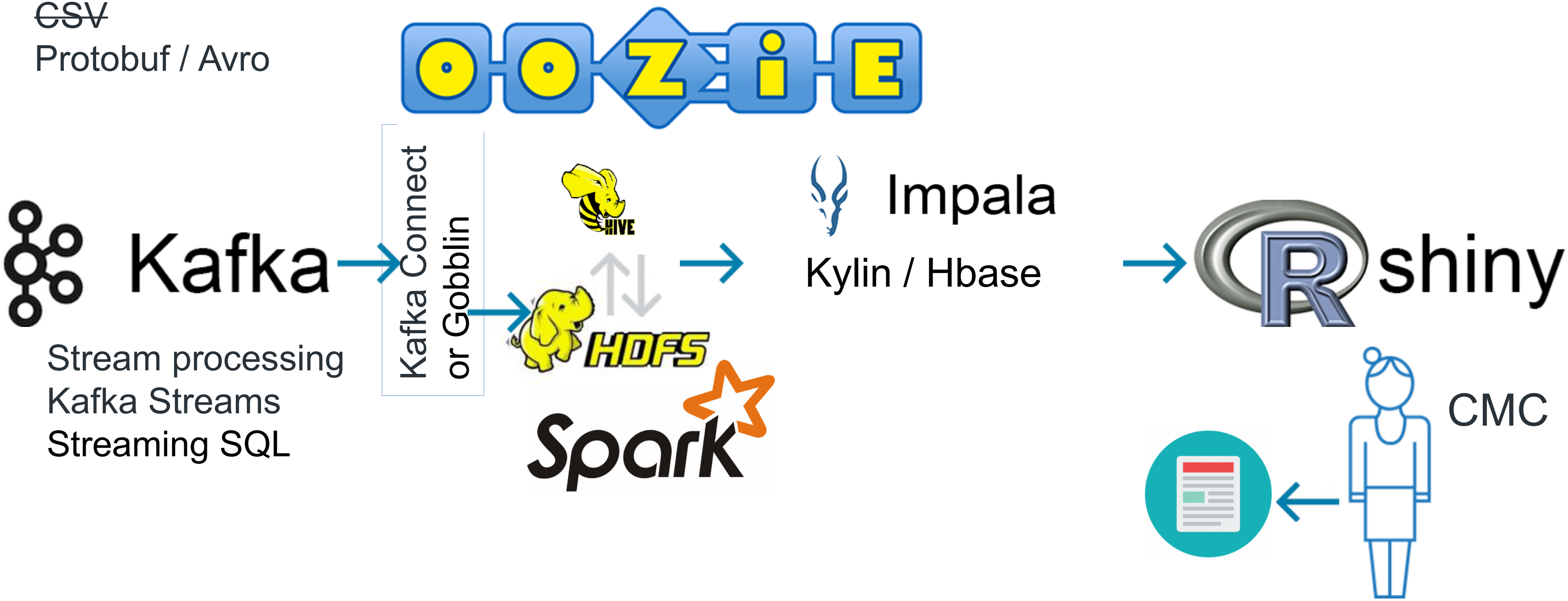
Message format:
~~CSV~~
Protobuf / Avro



Stream processing
Kafka Streams
Streaming SQL

Future data pipeline 2016/2017

Message format:
~~CSV~~
Protobuf / Avro



Kafka Connect or Gobblin

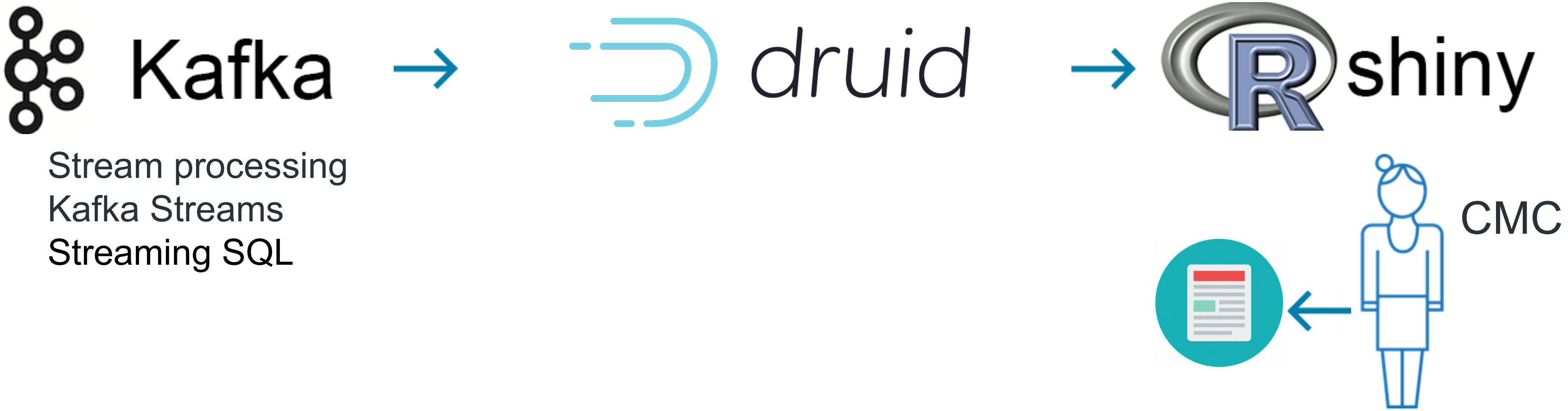
Impala
Kylin / Hbase

CMC

Stream processing
Kafka Streams
Streaming SQL

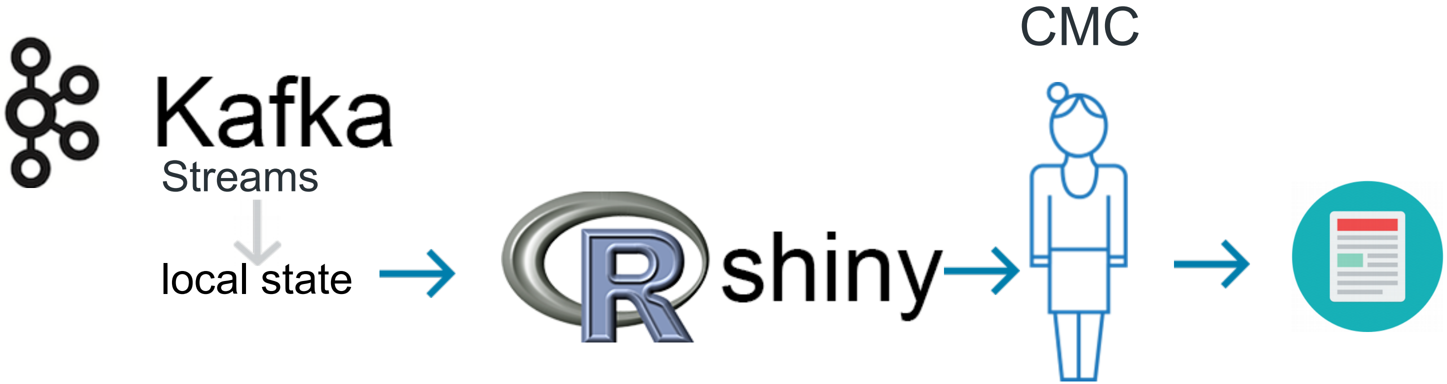
Future data pipeline 2016/2017

Message format:
~~CSV~~
Protobuf / Avro



Stream processing
Kafka Streams
Streaming SQL

Future data pipeline 2016/2017



Kafka Streams

local state

shiny

CMC

* https://www.confluent.io/blog/unifying-stream-processing-and-interactive-queries-in-apache-kafka/

Key challenges and learnings

Mastering hadoop

- Finding your log files
- Interpreting error messages correctly
- Understanding settings and how to use them to solve problem
- Store data in wide, denormalised Hive tables in parquet format and nested data types

Key challenges and learnings

Mastering hadoop

- Finding your log files
- Interpreting error messages correctly
- Understanding settings and how to use them to solve problem
- Store data in wide, denormalised Hive tables in parquet format and nested data types

Using hadoop

- Offer easy hadoop access to users (Impala / Hive JDBC with visualisation tools)
- Educate users on how to write good code, strict guidelines and code review
- deployment process: jenkins deploys git repository with oozie definitions and hive scripts to hdfs

Key challenges and learnings

Mastering hadoop

- Finding your log files
- Interpreting error messages correctly
- Understanding settings and how to use them to solve problem
- Store data in wide, denormalised Hive tables in parquet format and nested data types

Using hadoop

- Offer easy hadoop access to users (Impala / Hive JDBC with visualisation tools)
- Educate users on how to write good code, strict guidelines and code review
- deployment process: jenkins deploys git repository with oozie definitions and hive scripts to hdfs

Bad parts

- HUE (the standard GUI)
- Write oozie workflows and coordinators in xml, not through the Hue interface
- Monitoring impala
- Still some hard to find bugs in Hive & Impala
- Memory leaks with Impala & Hue: Failed queries are not always closed properly

Thank you!

Questions and comments?

Clemens Valiente

Senior Data Engineer
trivago Düsseldorf

Originally a mathematician
Studied at Uni Erlangen
At trivago for 5 years

Email: clemens.valiente@trivago.com
in de.linkedin.com/in/clemensvalien

Resources

- Gobblin: https://github.com/linkedin/gobblin
- Impala connector for dplyr: https://github.com/piersharding/dplyrimpaladb
- Querying Kafka Stream's local state: https://www.confluent.io/blog/unifying-stream-processing-and-interactive-queries-in-apache-kafka/
- Hive on Spark: https://cwiki.apache.org/confluence/display/Hive/Hive+on+Spark %3A+Getting+Started
- Parquet: https://parquet.apache.org/documentation/latest/
- ProtoBuf: https://developers.google.com/protocol-buffers/

Thanks to Jan Filipiak for his brainpower behind most projects, giving me the opportunity to present them