devoteam

Data foundation & pipelining

Anders Kring / 27 / 03 / 2020.

Innovative technology consulting for business.

DevoTalks

En webinar serie af digitale indspark

VELKOMMEN

DIGITALE
indspark

FORSKELLIGE
emner

SKIFTENDE
oplægsholdere

20 + 10
minutter

Lidt praktik inden start

Har du spørgsmål
kan du stille dem i chat-funktionen, og de vil blive besvaret efter 'talk'en'

Vil du gerne gense slides
bliver denne DevoTalk optaget og gjort tilgængelig på sitet devotalks.dk

Lidt om mig

-og den mulige grund til at jeg sidder her

- **Anders Kring**

  - data strateg & engineer.

- **Jeg arbejder primært med kunder der er i starten af deres datarejse, eller kunder der ønsker at blive mere datadrevne.**

  - er vild med når man kan gøre komplekse ting simple.

- **Udvikling af data strategier, governance og data fundamentet.**

  - det nogle vil kalde alt det u-sexede i data-analyse-fagligheden.

- **Udvikling af prototyper og demoer på teknologier.**

  - så heldig er jeg.

Agenda

Vi skal snakke om hvad data foundation & pipelining er.

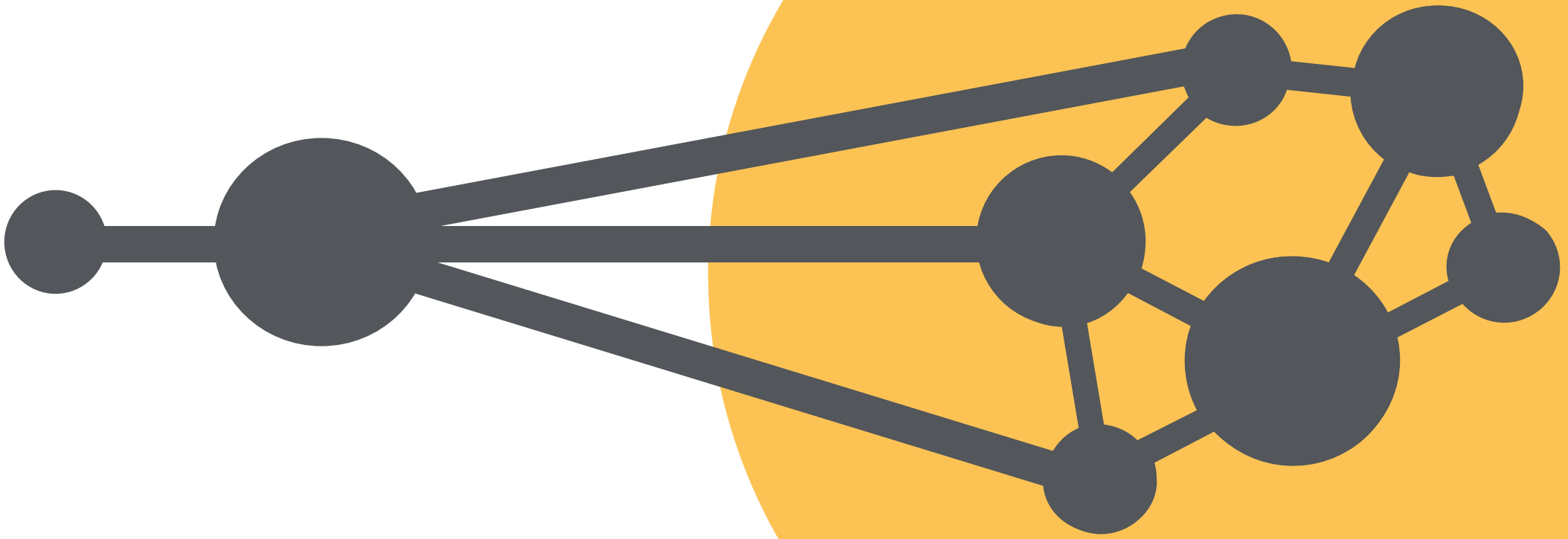
- Og hvorfor det er så vigtigt

Kigge på forskellige eksempler på implementeringer af datapipelines.

- Hvis vi når det

Hvad snakker du om ?

-en smule om begreber



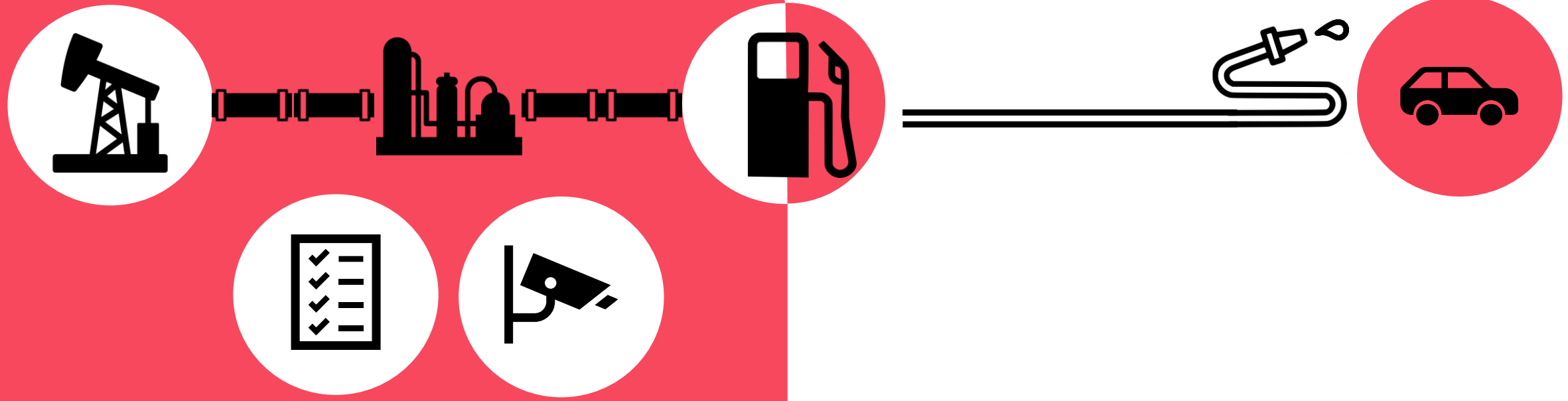
Data Analyse Platform

Data Integrations Platform

Data Foundation er roden til … alt godt …

Data Foundation

Data Science

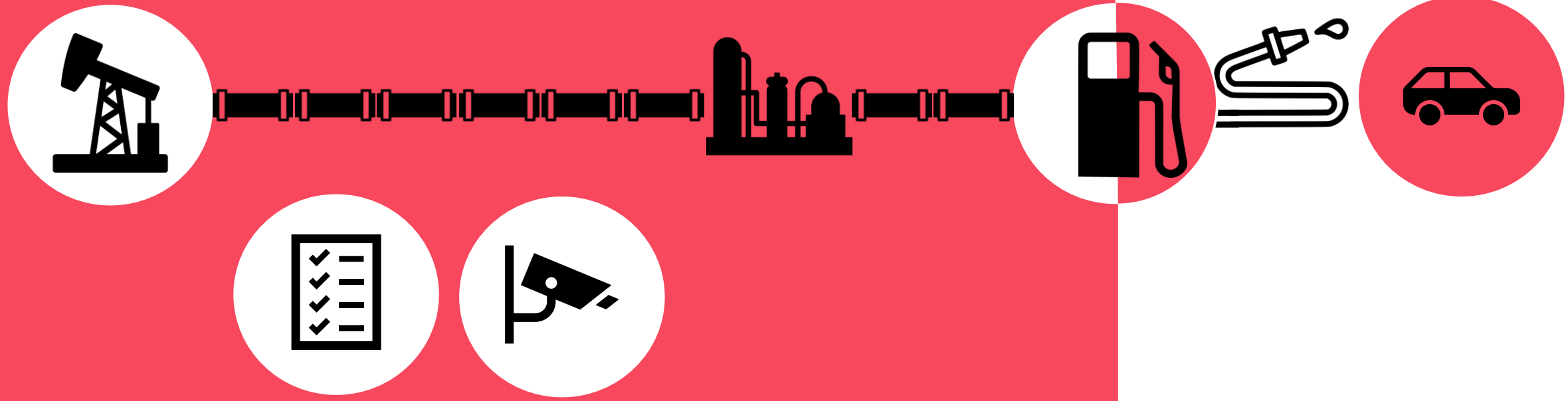


Data Foundation er roden til … alt godt …

Data Foundation

80 %

Data Science



En typisk data pipeline proces … er mangfoldig

Kilden

- Identifikation af kilde.
- Identifikation af kildeejer.
- Afklaring af hvordan denne kan tilgås, og hvor ofte.
- SLA for kilde.
- Implementation af data høster.

Rensningen / Transformationen

- Identificere og udbedre fejlmålinger.
- Identificere huller i data.
- Normalisere data.
- Digitalisere analog data.
- Strukturere data.
- Pseudonymisering / Anonymisering.
- Aggregering af data.
- Samkøring af data.

Opbevaringen

- Gemme ustruktureret data i et object storage.
- Gemme struktureret data en relationel database.
- Gemme dokumenter i et dokument database.
- Gemme streaming-data i specielt BigSql.

Tilgængeliggørelse

- Udvikle API'er der kan tilgås af eksterne applikationer.
- Oprette streaming services til live dashboards.
- Styre adgangskontrol til brugere og databaser.

Den dårlige gode nyhed er !

Der er utroligt mange måder at implementere data pipelines på…

Der er utroligt mange måder at implementere data pipelines på…

Typer af løsninger … kategorier

Kommerciel	Open Source	Build Self
Services der følger med valg af infrastruktur.	Hurtige forbedringer.	Frit valg af teknologi.
Vendor lock-in.	Bred mulighed for forskellige infrastrukturer.	Frit valg af infrastruktur.
Højt abstraktions niveau.	Relativt abstraktions niveau.	Lavt abstraktions niveau.
Lav konfigurerbarhed.	Relativt konfigurerbar.	Fuld konfigurerbar.
↓	↓	↓
Pentaho, Data Factory	NiFi, Luigi, Airflow	Python, Java, R

Hvad er kravene og ambitionerne til, med og fra din platform, din data aftagere og din udfordring.

Hvilken løsning skal man vælge ?

Kommerciel

Hvis du allerede har en strategi om leverandør.

Hvis du ikke har eller planlægger at have en data udviklings afdeling.

Hvis du vil kunne on-boarde nye medarbejdere hurtigt

Hvis du vil være platform uafhængig.

Hvis du vil udnytte on-prem hardware.

Hvis du har meget specifikke og eksotiske behov for din data-bearbejding

Open Source

Hvis du gerne vil have kontrol over hvornår du opgradere din software.

Hvis du ikke ønsker vendor lock-in.

Hvis du gerne vil bygge din arkitektur modulært

Gerne vil bruge generelle kendte standart systemer.

Du gerne vil udnytte den professional services der normalt kommer med et kommercielt produkt

Hvis garantier og SLA'er er kritiske for din virksomhed.

Build Self

Hvis du vil have fuld kontrol over alle trin i din pipeline.

Hvis du har meget specielle cases for dine transformationer.

Hvis dine pipelines er meget simple.

Hvis du er bange for at miste dine talenter.

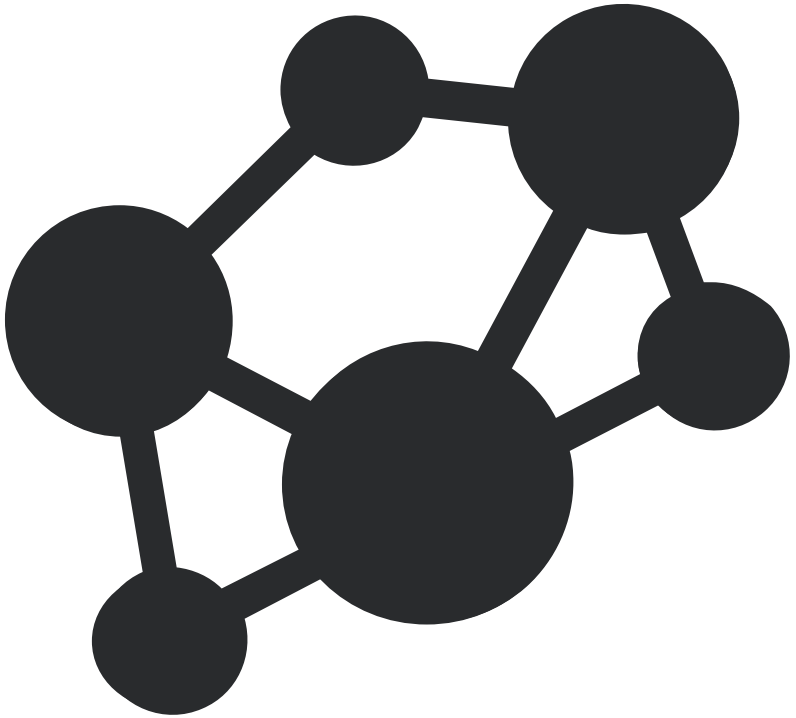
Hvis du ikke allerede har en udviklings afdeling og udviklings processer.

Hvis du ikke vil have et stort overhead på bug-fixing og optimering

Takeaways…

Det er aldrig klogt at vælge en platform baseret på "det vi plejer"

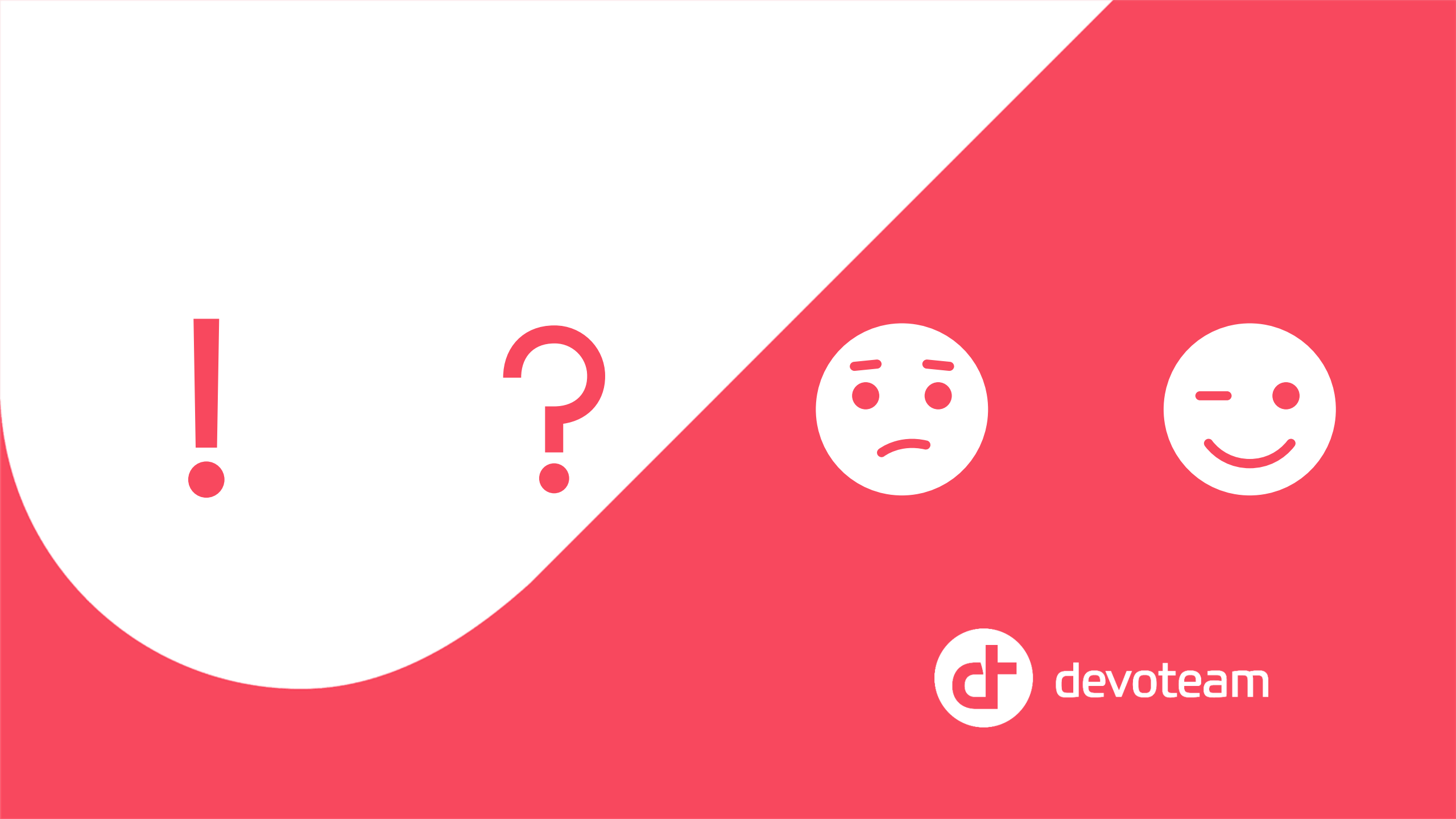
- Selv om det er trygt



Takeaways…

Det er ikke så svært at komme igang…

… selv om nogen måske prøver at bilde dig det ind


!
?
devoteam

Data pipeline eksempler

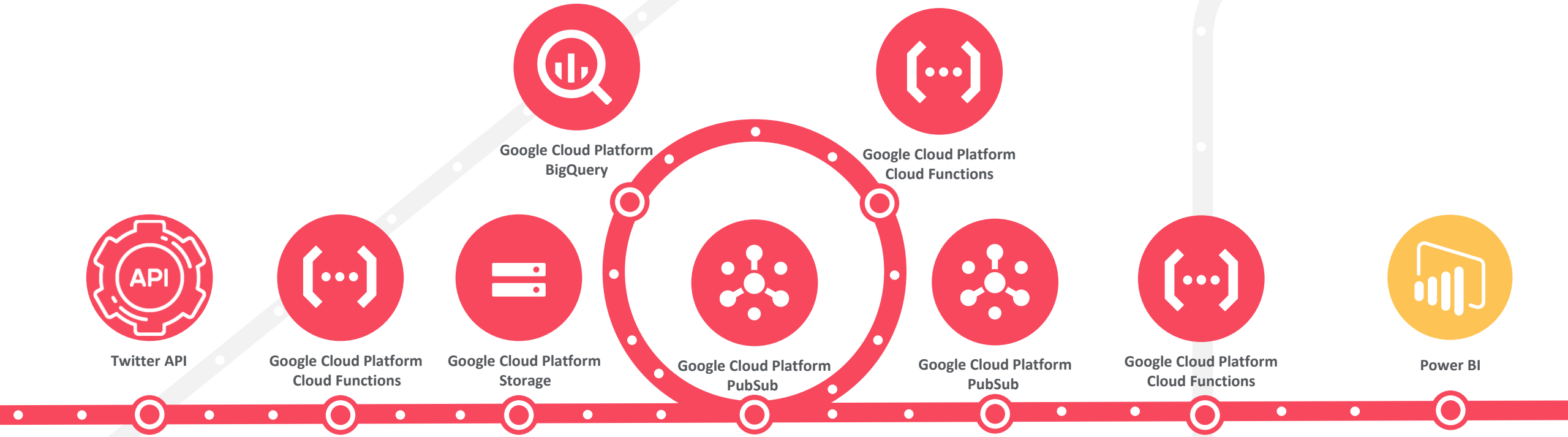
Vi ser hvor langt vi når...

Hvad snakker politikere og journalister om før et valg…

… på Twitter

- Løbende hente Twitter beskeder fra politikere, journalister og politiske partier i et valgår.
- Hent Tweets, og kontinuert lave nye analyser på dem.
- Uklart mål med data.
- Sikre ikke at nå API limit på twitter.
- Høstning og analyse er ikke tidskritisk.
- Vise data close-to-live.

Twitter analyse pipeline.



”Everywhere you go, always take the weather with you”

… fra OpenWeatherMap

- Gem det nuværende vejr i en database til senere analyse.
- Gør data tilgængelig til visuel analyse live.
- Brug platform agnostiske services.
- Minimere programmering.

Vejr data pipeline.

Apache
NiFi

MySql

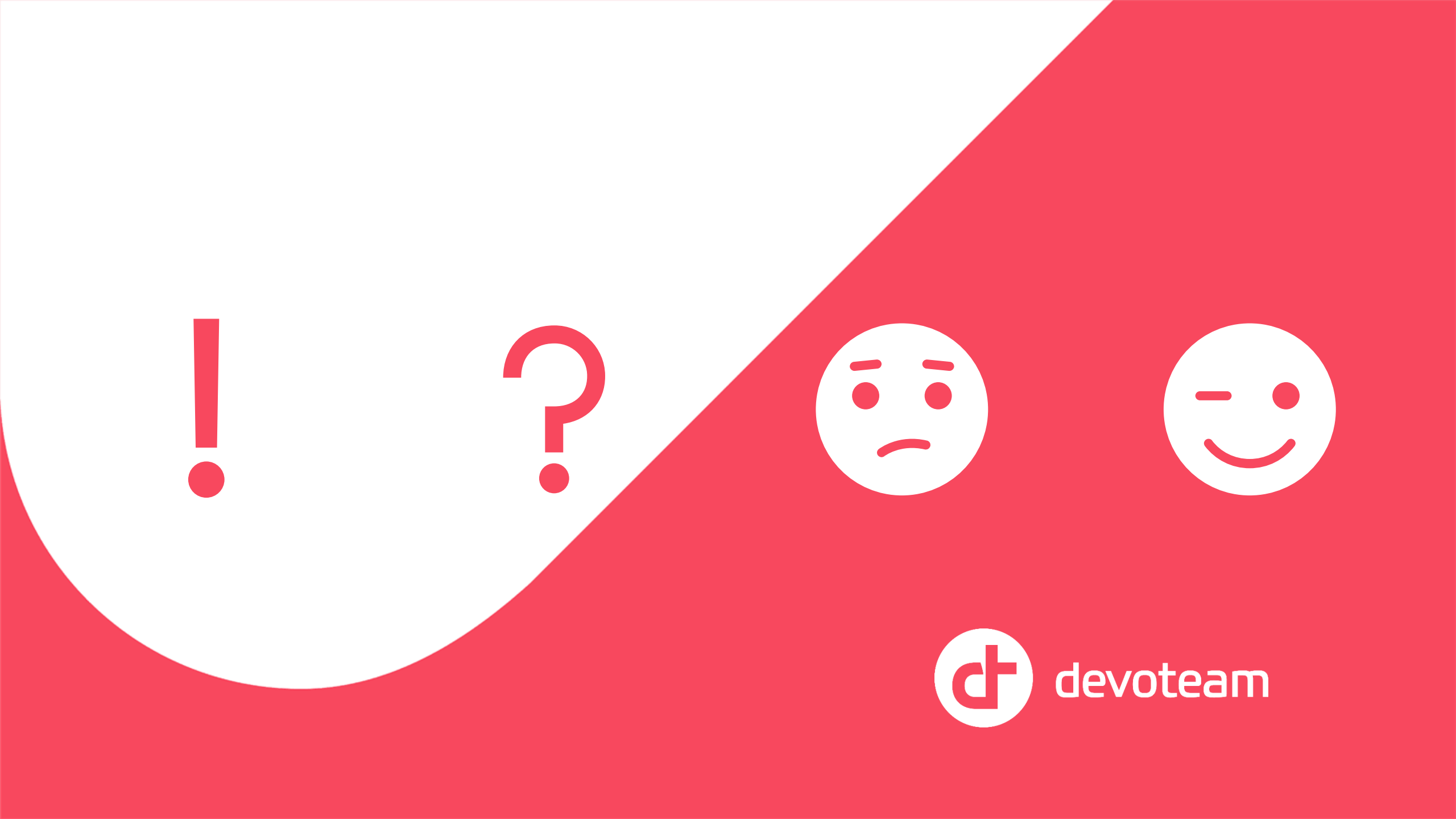
Apache
SuperSet

Det kunne være godt at vide om der er parkering til mig i næste uge..

… Data Science med IOT

- Gøre parkerings-data tilgængeligt til at lave data-science på.
- Gøre det muligt at finde ud af om der er parkeringspladser på kontoret.
- Hvor billigt kan man gøre det ?

Parkeringssensor pipeline.


!
?
devoteam