BigData et traitement de données à l'échelle

MAXIME LEMAIRE – OLIVIER POTTIÉ – TIMELAB

Innovativ Days
Smarter Together

Un peu d'histoire...

En 2006, Google publie le whitepaper
"*Bigtable: A Distributed Storage System for Structured Data*"

voit le jour quelques années plus tard et confirme cette nouvelle tendance


MEGADONNÉES !!

Avènement d'une nouvelle technologie :

Le NoSQL

pour

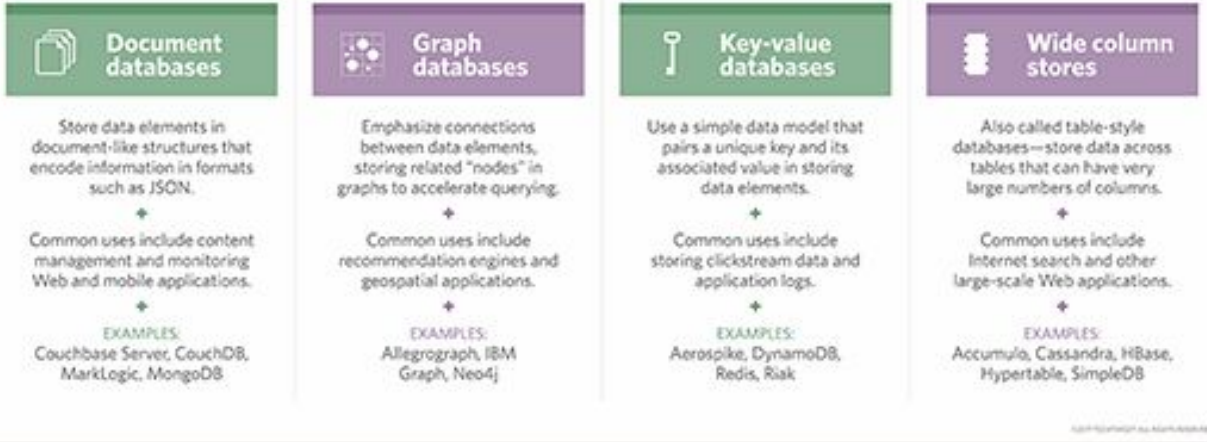
Stocker de la donnée en masse à la Google

merci l'Open Source encore une fois quand même...

Bon en vrai ça ressemble à ça...

All in the NoSQL Family

NoSQL databases are geared toward managing large sets of varied and frequently updated data, often in distributed systems or the cloud. They avoid the rigid schemas associated with relational databases. But the architectures themselves vary and are separated into four primary classifications, although types are blending over time.



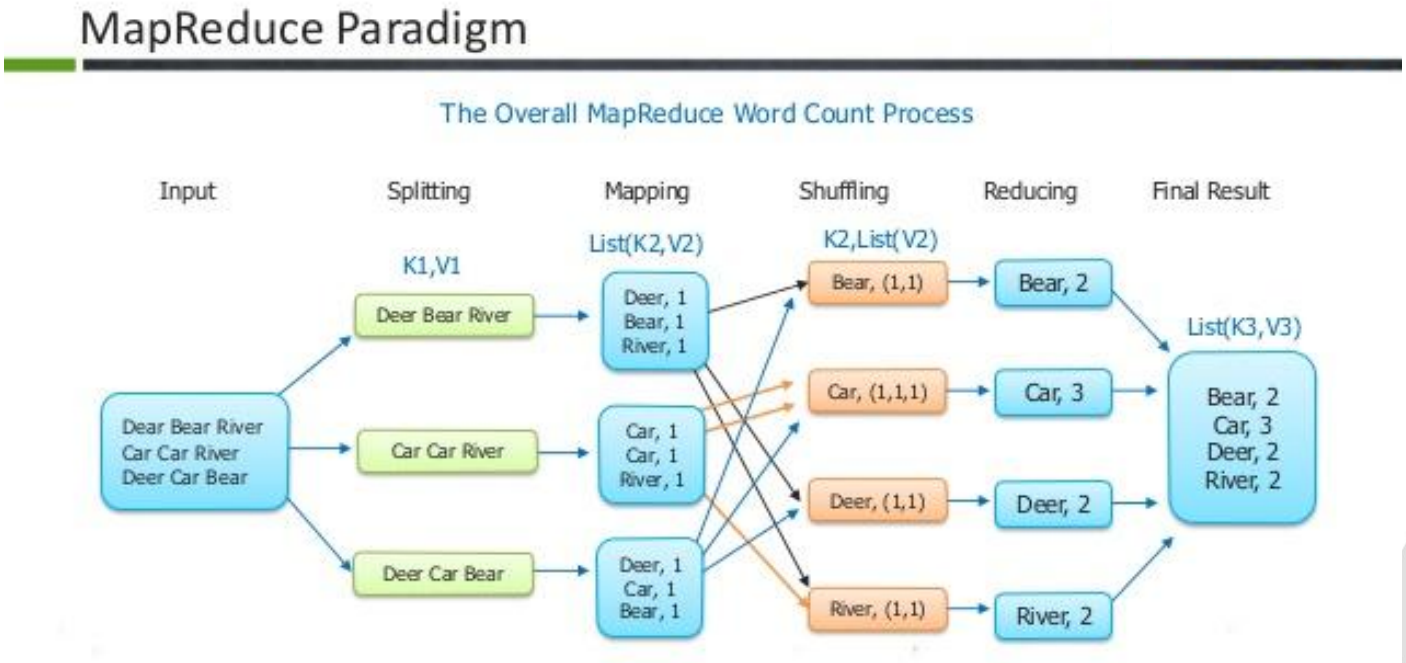
Innovativ Days
Smarter Together

Particularités

- Différents types de datastore en fonction du besoin
- Pas de schéma ou de structure de données à priori
- Stockage distribué / répliqué
- Scalabilité horizontale
- Tolérance à la panne et haute disponibilité
- ACID vs BASE
  Atomicity, **C**onsistency, **I**solation, **D**urability VS **B**asically **A**vailable, **S**oft state, **E**ventual consistency

Innovativ Days
Smarter Together

Et le traitement ?



Innovativ Days
Smarter Together

Et le traitement ?

Le MapReduce de Hadoop c'est bien mais :

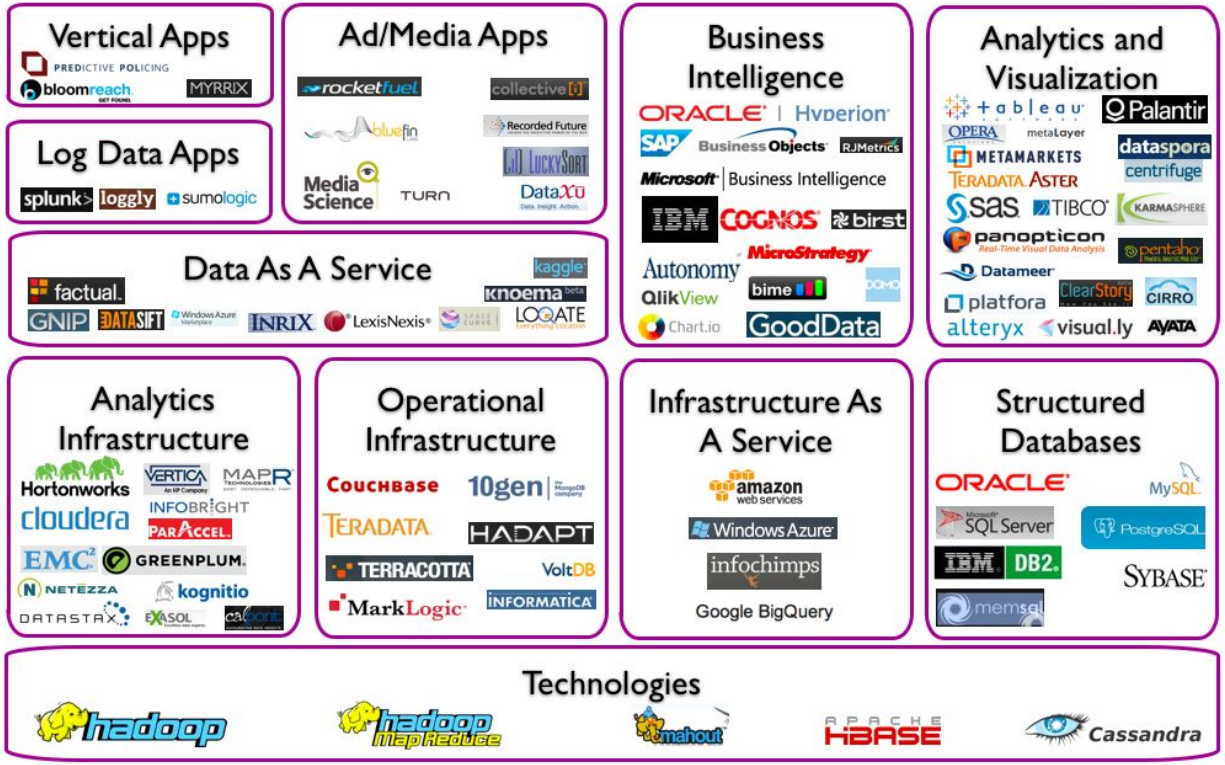
- c'est lent
- assez gourmand en ressources
- plutôt adapté au traitement batch et pas du tout au temps-réel

De nouveaux frameworks apparaissent, le plus notable 

Le machine learning devient également une composante importante de ces nouveaux frameworks : il faut créer de la valeur !

Innovativ Days
Smarter Together

BIG DATA LANDSCAPE 2012

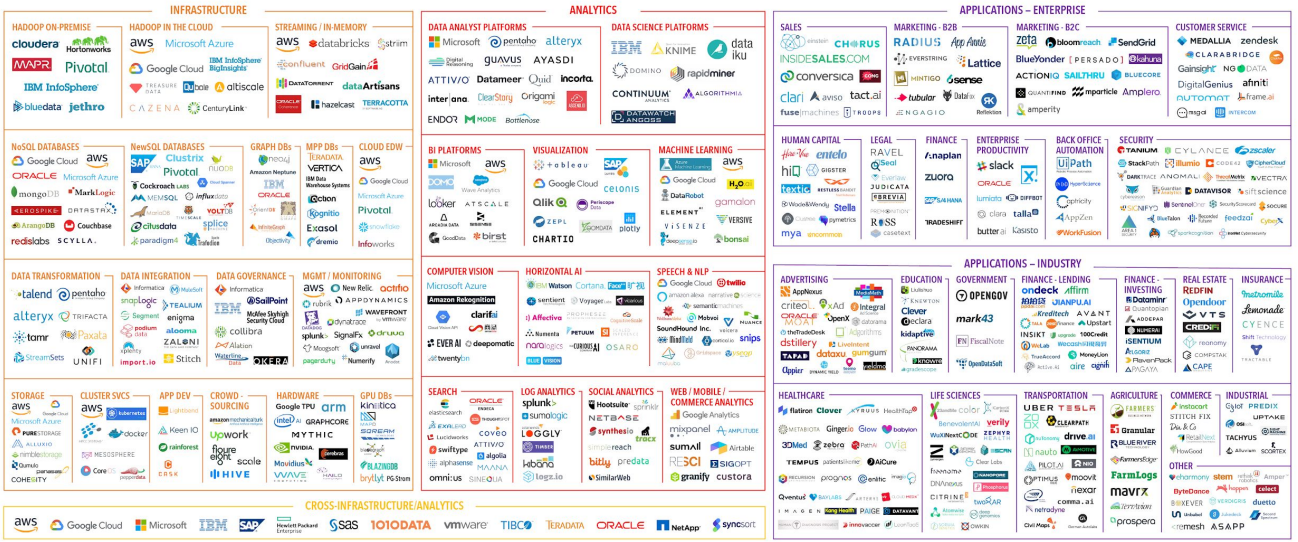


dave@vcdave.com

blogs.forbes.com/davefeinleib

Innovativ Days
Smarter Together

BIG DATA (& AI) LANDSCAPE 2018



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Innovativ Days
Smarter Together

En parallèle de cette évolution de la gestion de la donnée et de l'utilisation des ressources, la gestion de la couche applicative évolue également grâce aux containers et plus récemment aux orchestrateurs.

Innovativ Days
Smarter Together

Chronologie de l’évolution de notre architecture

Innovativ Days
Smarter Together

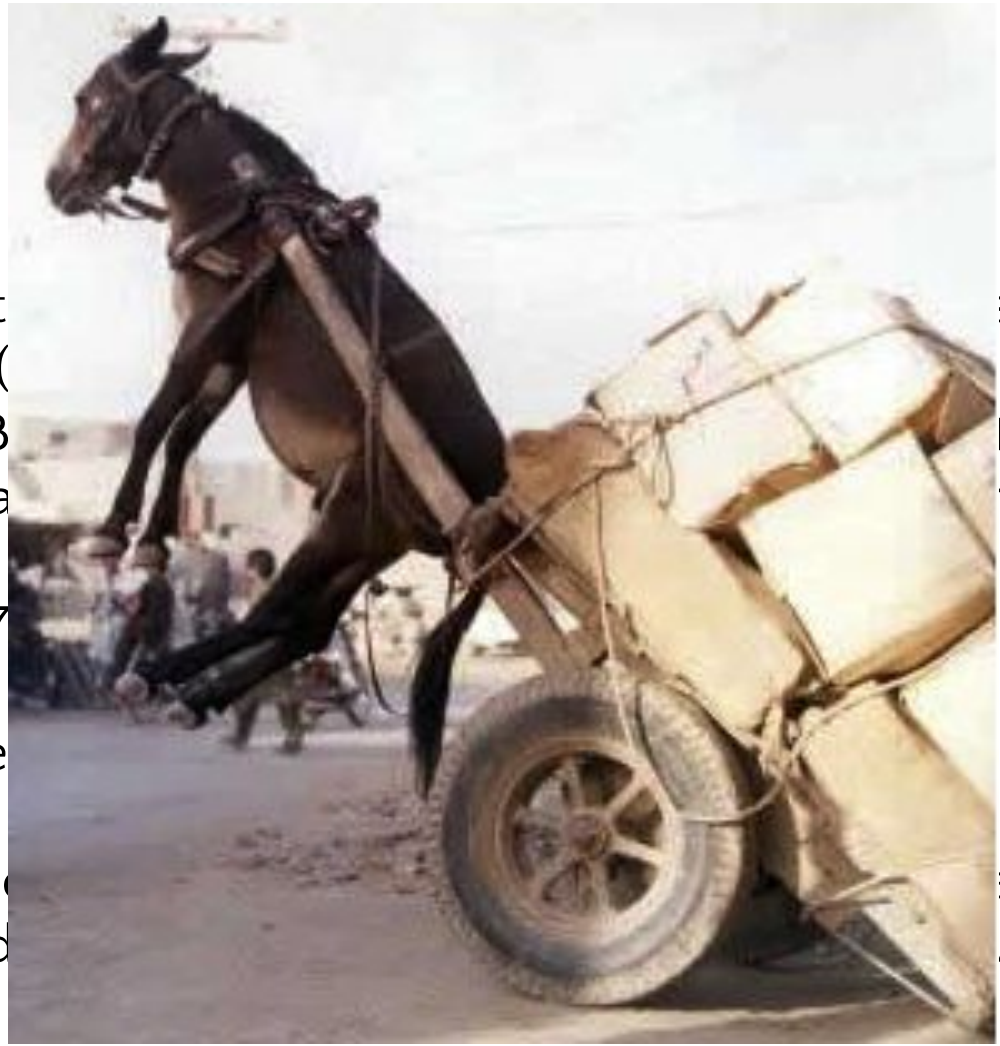
2013-2014

On teste toutes les technos qui nous semblent intéressantes :
Hadoop :(
MongoDB puis ElasticSearch pour les données utilisateurs
HBase, Cassandra, OpenTSDB, KairosDB pour les données de capteurs
Redis, Kafka (0.7 !) pour le messaging

On essuie les plâtres :

Les technos sont jeunes, pas toujours stables, complexes et on manque d'outil d'administration pour se simplifier la vie

Innovativ Days
Smarter Together



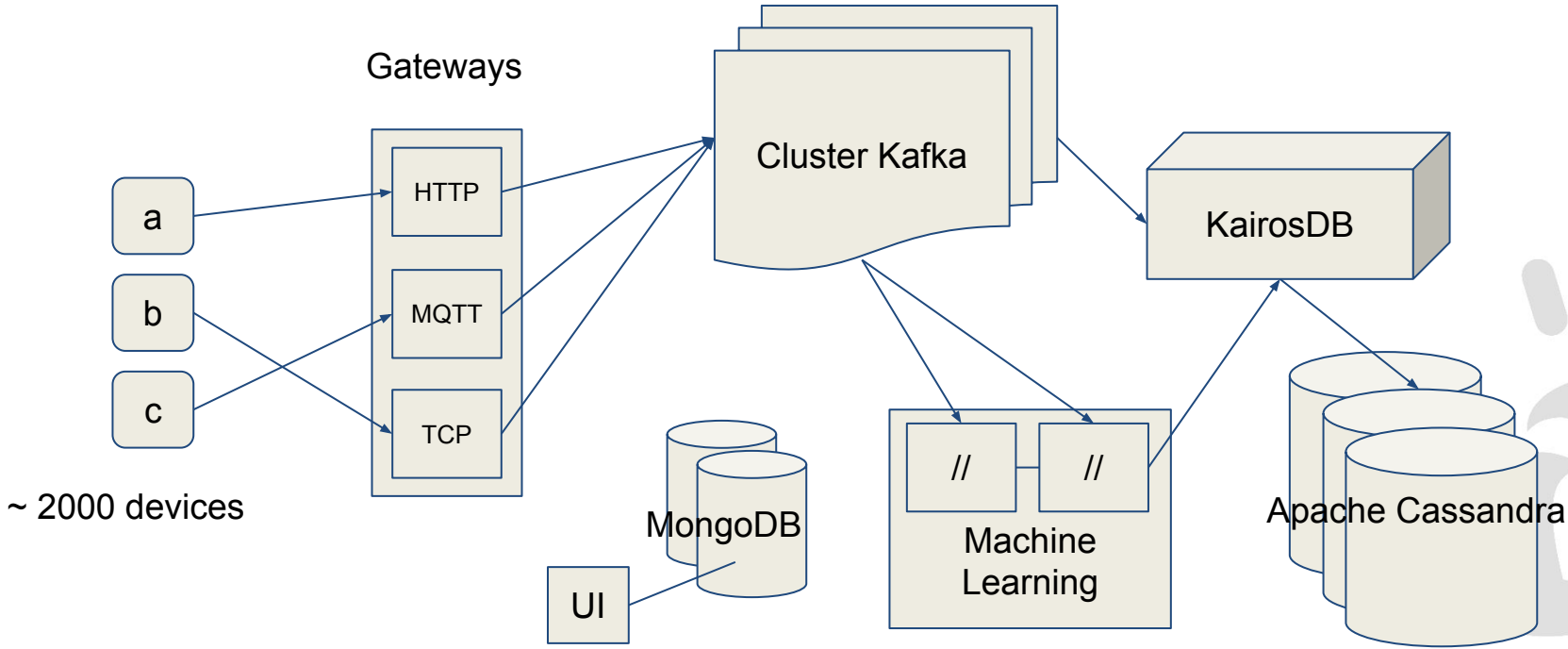
On teste t ssantes :
Hadoop :(
MongoDB teurs
HBase, Ca nnées de
capteurs
Kafka (0.7

On essuie

Les techn exes et on
manque d vie

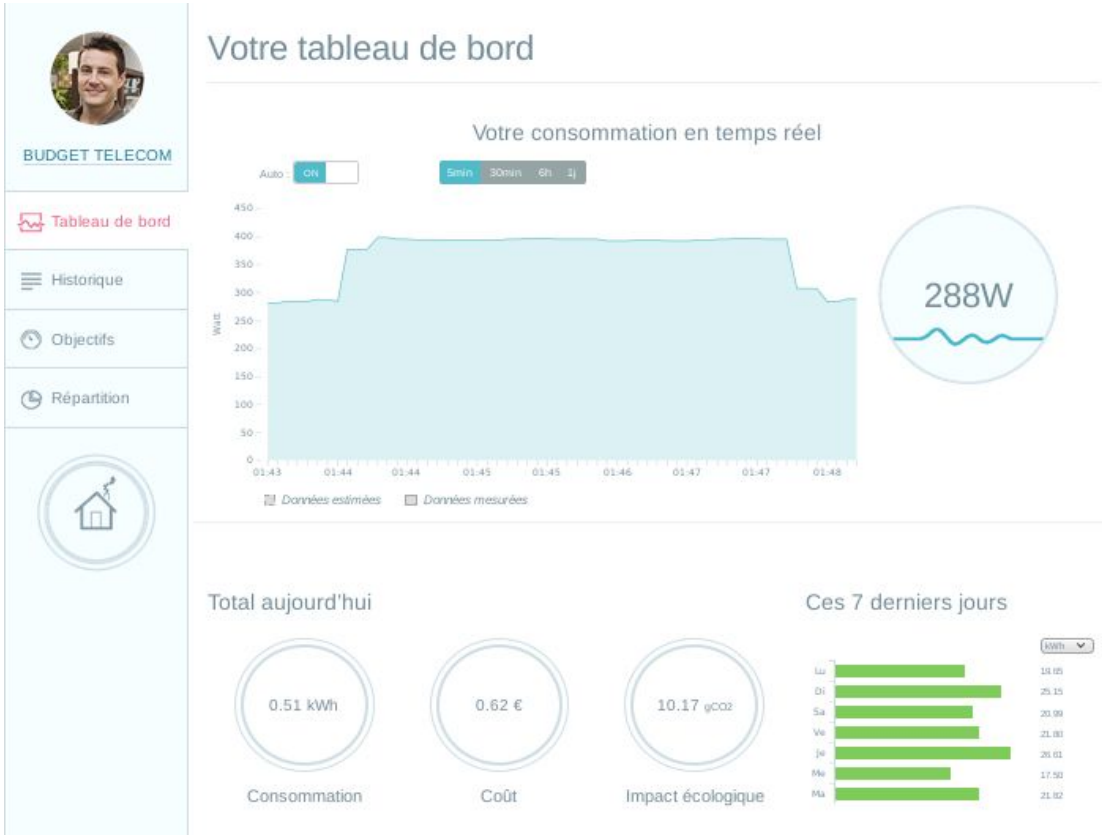
Innovativ Days
Smarter Together

2014-2015



On a quand même quelque chose qui tourne..

Innovativ Days
Smarter Together



ça marche et c'est performant !

Innovativ Days
Smarter Together

2015-2016

- On manage Cassandra grâce à DataStax Enterprise
- On passe progressivement la totalité des applications sous Docker
- On migre nos serveurs dédiés chez Google Cloud et on profite de la virtualisation et d'une vraie scalabilité horizontale
- On orchestre la totalité de la plateforme (hors Cassandra) sous Kubernetes

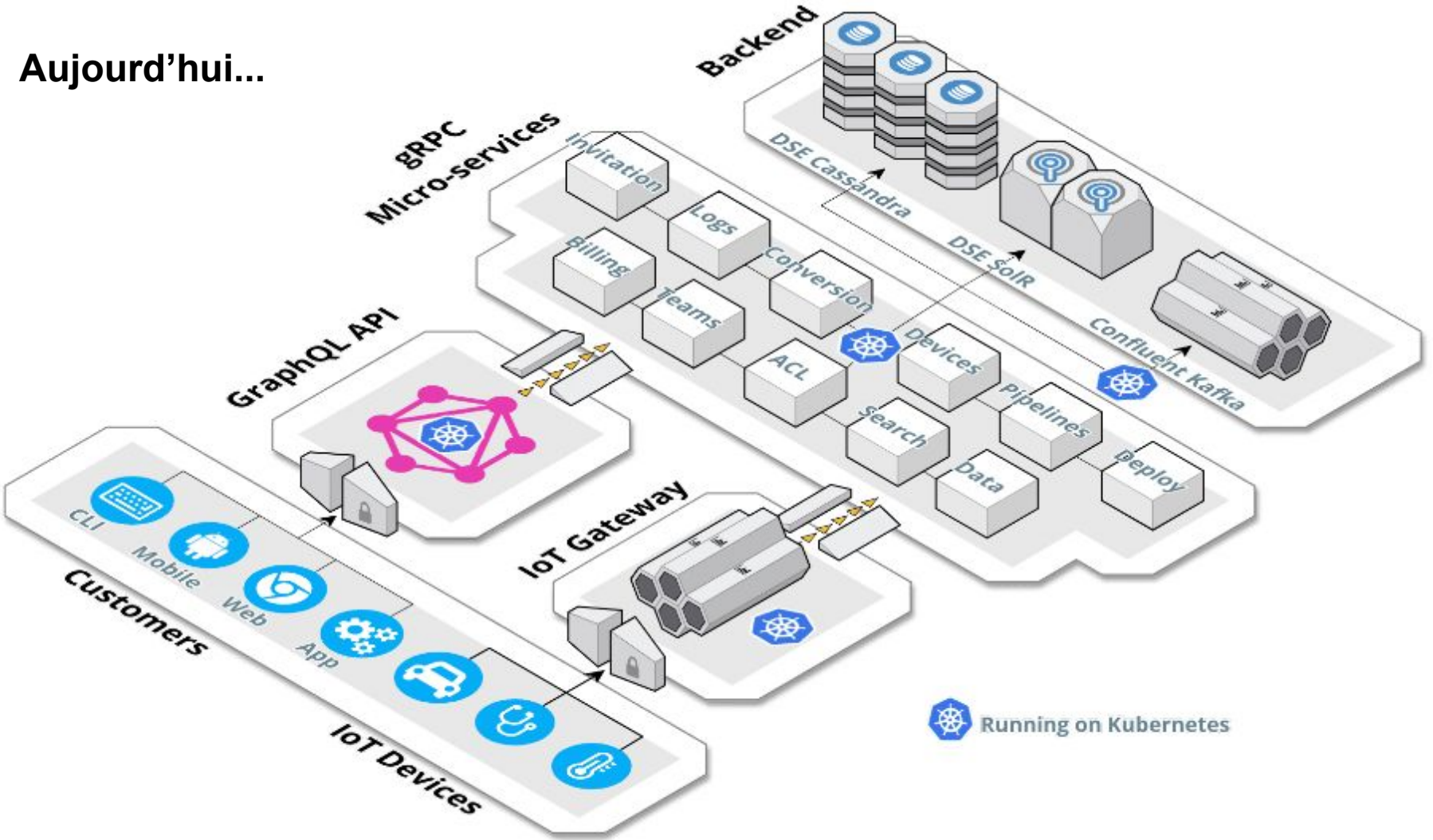
...on essuie d'autres plâtres ;)

Innovativ
Days
Smarter Together

- On ma
- On pa er
- On mi e la virtual
- On or ubernetes

tres plâtres ;)

Aujourd'hui...


Backend
gRPC
Micro-services
Invitation
Logs
Billing
Conversion
Teams
ACL
Devices
Search
Pipelines
Data
Deploy
DSE Cassandra
DSE SolR
Confluent Kafka
GraphQL API
IoT Gateway
CLI
Mobile
Web
App
Customers
IoT Devices
Running on Kubernetes

Innovativ Days
Smarter Together

Pour essuyer moins de plâtres..

- Importance de définir un workload cible (batch, temps-réel, r, w, rw)
- Bien choisir le type de base de données NoSQL qui répond au besoin
- Comprendre les implications du modèle de consensus utilisé (PAXOS, Gossip, ...)
- Monitorer monitorer monitorer
- Optimiser très tôt la configuration, mesurer l'impact, REPEAT
- Développer toujours vos applications en prenant en compte que le réseau n'est pas fiable. Préférer les applications Stateless / Crash-only.

Innovativ Days
Smarter Together

Merci de votre attention !
Des questions ?

timelab