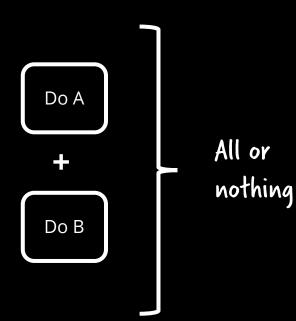
Lost in transaction?

Strategies to deal with (in)consistency in distributed systems







once upon a time:

```
try {
  tx.begin();
  doA();
  doB();
  tx.commit();
} catch (Exception e) {
  tx.rollback();
}
```

or simply:

```
@Transactional
public void createCustomer(Customer cust) {
   // ...
}
```

Atomicity (onsistency solation Durability

Distributed systems

Distributed systems



Article Talk

WIKIPEDIA The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article Donate to Wikipedia

Wikipedia store

Interaction

Help

About Wikipedia Community portal Recent changes

Contact page

Tools

What links here Related changes Upload file

Fallacies of distributed computing

From Wikipedia, the free encyclopedia

The fallacies of distributed computing are a set of assertions made by L Peter Deutsch and others at Sun Microsyste

Contents [hide]

1 The fallacies

2 The effects of the fallacies

3 History

4 See also

5 References

6 External links

The fallacies [edit]

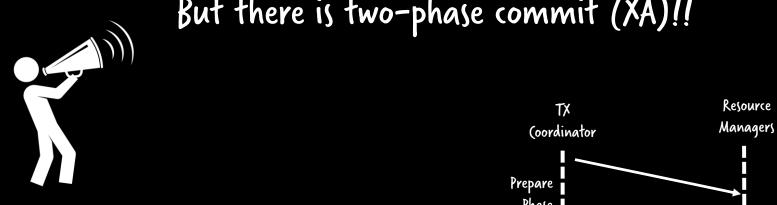
The fallacies are:[1]

1. The network is reliable.

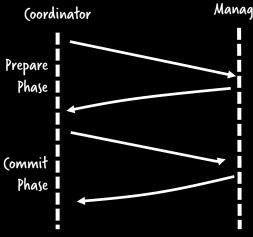
2. Latency is zero.

3. Bandwidth is infinite.

Distributed systems



But there is two-phase commit (XA)!!





Pat Helland

Distributed Systems Guru Worked at Amazon, Microsoft & Salesforce

Life beyond Distributed Transactions: an Apostate's Opinion

Position Paper

Pat Helland

Amazon.Com 705 Fifth Ave South Seattle, WA 98104 USA PHelland@Amazon.com

The positions expressed in this paper are personal opinions and do not in any way reflect the positions of my employer Amazon.com.

ABSTRACT

Many decades of work have been invested in the area of distributed transactions including protocols such as 2PC, Paxos, and various approaches to quorum. These protocols provide the application programmer a façade of global serializability. Personally, I have invested a nontrivial portion of my career as a strong advocate for the implementation and use of platforms Instead, applications are built using different techniques which do not provide the same transactional guarantees but still meet the needs of their businesses.

This paper explores and names some of the practical approaches used in the implementations of large-scale mission-critical applications in a world which rejects distributed transactions. We discuss the management of fine-grained pieces of application data which may be repartitioned over time as the application grows. We also discuss the design patterns used in sending messages between these repartitionable pieces of data



"Grown-Ups Don't Use Distributed Transactions

Pat Helland

Distributed Systems Guru Worked at Amazon, Microsoft & Salesforce

Starbucks does not use two phase commit

https://www.enterpriseintegrationpatterns.com/ramblings/18_starbucks.html

See. 4

Photo by John Ingle

Eric Brewer

But we forfeit "C" and "I" for availability, graceful degradation, and performance

This tradeoff is fundamental.

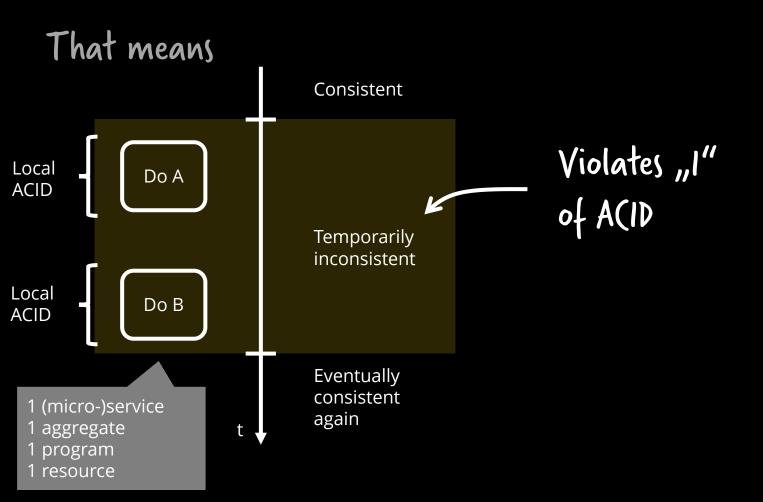
BASE:

- Basically Available
- Soft-state
- Eventual consistency

PODC Keynote, July 19, 2000

http://pld.cs.luc.edu/courses/353/spr11/notes/brewer_keynote.pdf

Atomicity (onsistency Isolation Durability



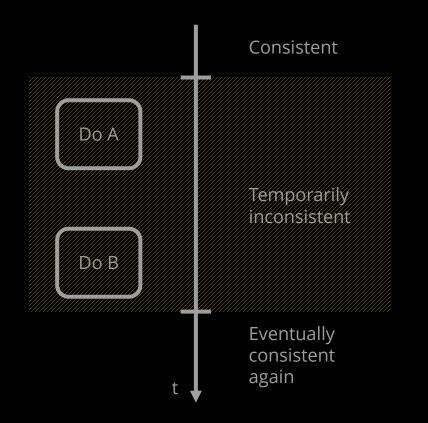


Photo by <u>Gerhard51</u>, available under <u>Creative Commons CC0 1.0 license</u>.





You



Pat Helland

"Building on Quicksand" Paper



D

A



Pat Helland

"Building on Quicksand" Paper

Associative (ommutative Idempotent Distributed 2.0

(a + b) + c = a + (b + c)

a + b = b + a

 $f(x) = f(\overline{f(x)})$



Photo by <u>pixabay</u>, available under <u>Creative Commons CC0 1.0 license</u>.

Requirement: Idempotency of services!



Photo by pixabay, available under Creative Commons CC0 1.0 license.

Requirement: Idempotency of services!

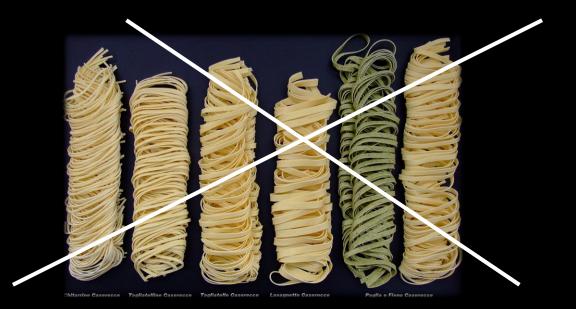
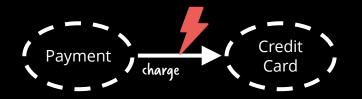
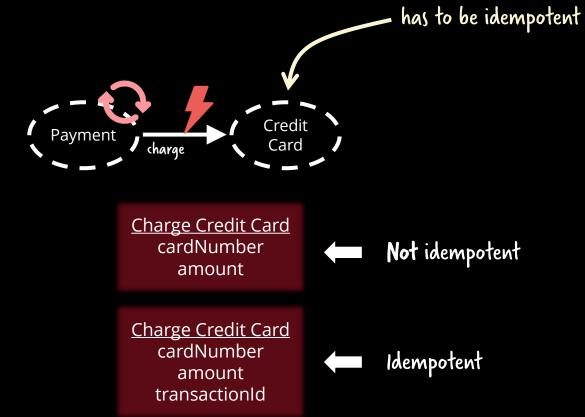


Photo by <u>Chr.Späth</u>, available under <u>Public Domain</u>.

Example



Strategy: retry

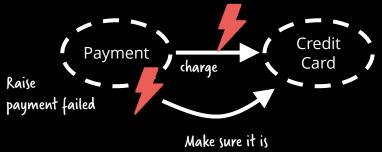


Distributed

It is impossible to differentiate certain failure scenarios: (lient Service Provider X Y Y Y

Independant of communication style!

Strategy: (leanup



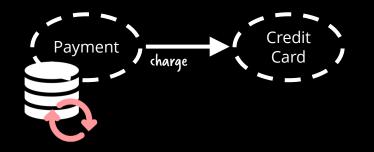
not charged!

Cancel charge cardNumber amount transactionId

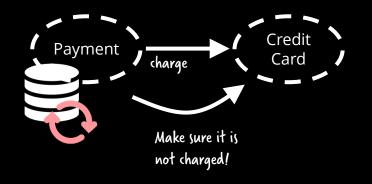


Some communication challenges require state.

Strategy: Stateful retry



Strategy: Stateful retry



Warning: Contains Opinion



Bernd Ruecker

(o-founder and (hief Technologist of (amunda

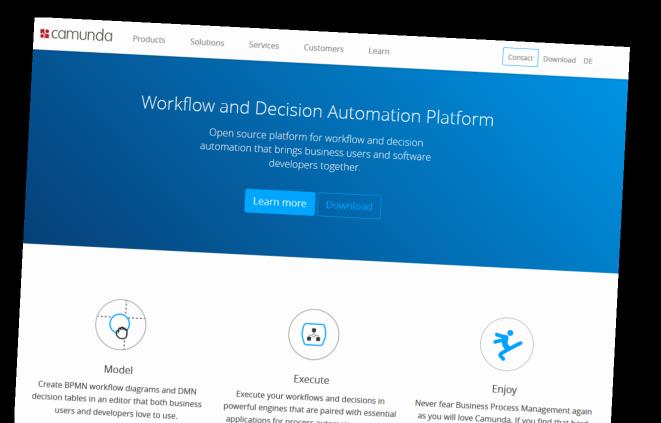


Berlin, Germany

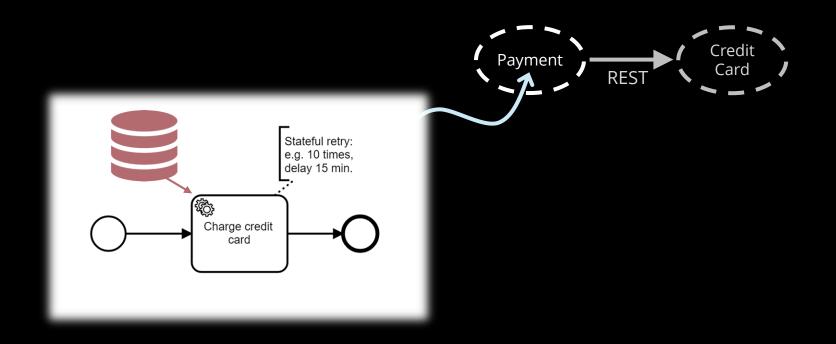


bernd.ruecker@camunda.com
@berndruecker

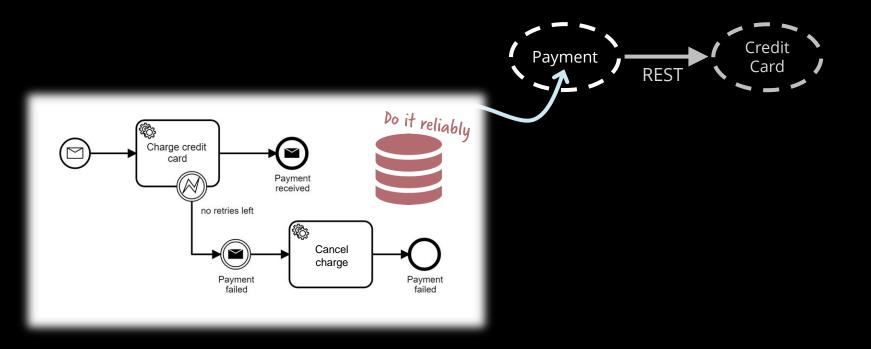
Let's use a lightweight OSS workflow engine for this:



Stateful retry



Stateful retry & cleanup



Live hacking

https://github.com/flowing/flowing-retail/tree/master/rest

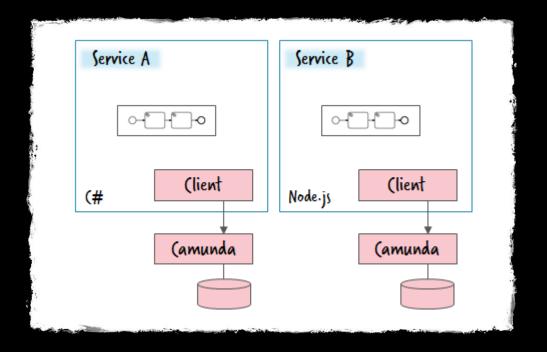


Embedded Engine Example (Java)

pring Boot	Spring Boot
Infrastructure (ockpit (amunda	(ockpit (amunda
omain	
Service A	Service B

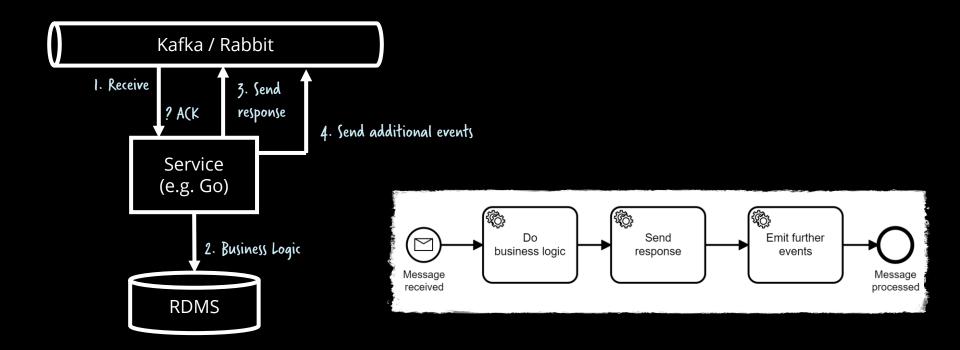
https://blog.bernd-ruecker.com/architecture-options-to-run-a-workflow-engine-6c2419902d91

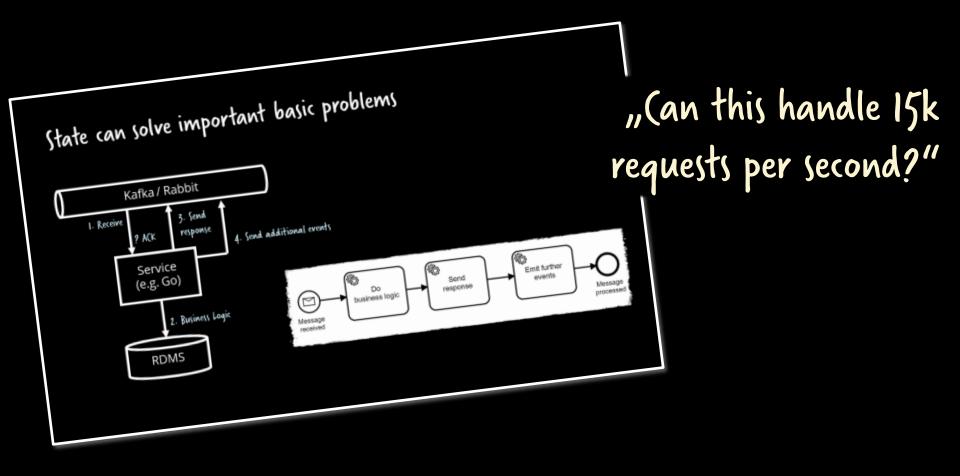
Remote Engine Example (Polyglot)

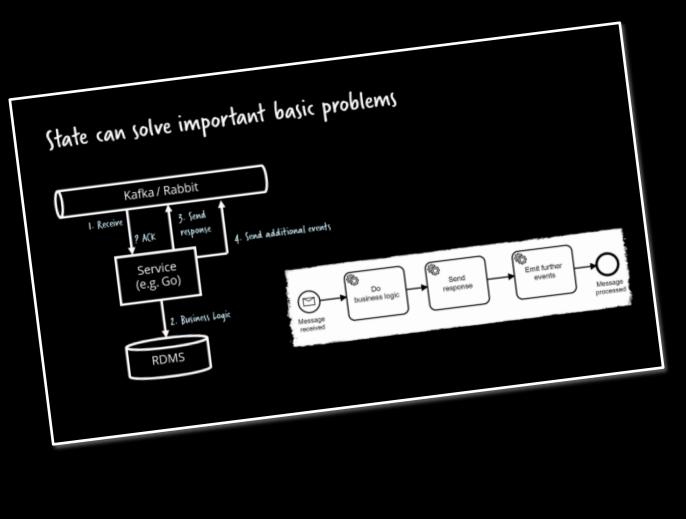


https://blog.bernd-ruecker.com/architecture-options-to-run-a-workflow-engine-6c2419902d91

A relatively common pattern





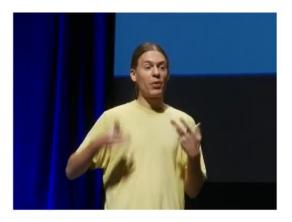


Zeebe

"Yes."





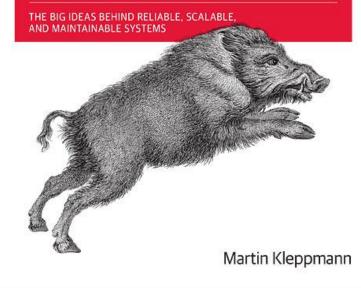


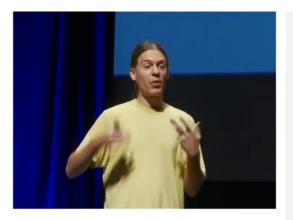


Sept 25-26, 2015 thestrangeloop.com

O'REILLY"

Designing Data-Intensive Applications







Sept 25-26, 2015 thestrangeloop.com

Without cross-service transactions: Compensating transactions) ~ abort/rollback at app level (Garcia-Molina & Salem, 1987) Apploaies C (Helland & Campbell, 2009) after the fact, rather than proventing them)



PatHelland's WebLog

Memories, Guesses, and Apologies

Rate this article *****



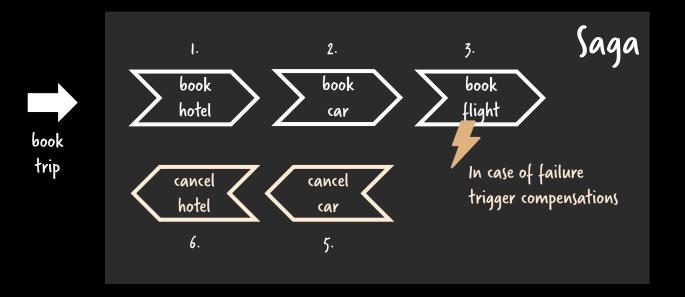


Well, here I am blogging on the bus with my newly installed Windows Live Writer!!! This blog is a text version of a five minute "Gong Show" presentation I did at CIDR (Conference on Innovative

5

All <u>https://blogs.msdn.microsoft.com/pathelland/2007/05/15/memories-guesses-and-apologies/</u> nion about

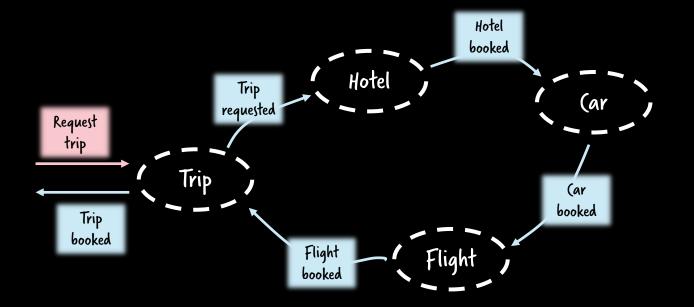
(ompensation - the classical example



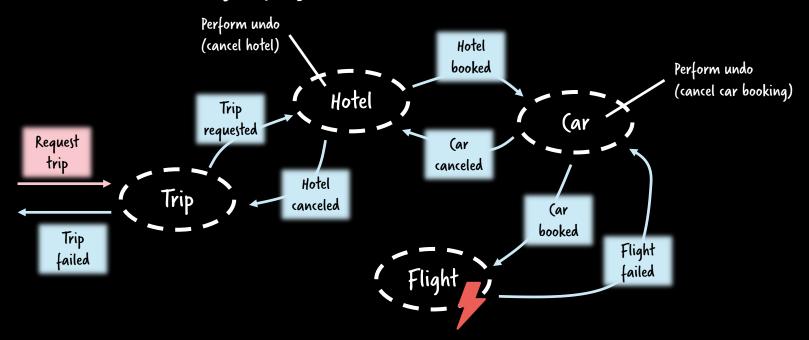
2 alterntive approaches: choreography & orchestration

A Contraction of the second se

Event-driven choreography



Event-driven choreography





The danger is that it's very easy to make nicely decoupled systems with event notification, without realizing that you're losing sight of that larger-scale flow, and thus set yourself up for trouble in future years.

https://martinfowler.com/articles/201701-event-driven.html



The danger is that it's very easy to make nicely decoupled systems with event notification, without realizing that you're losing sight of that larger-scale flow, and thus set yourself up for trouble in future years.

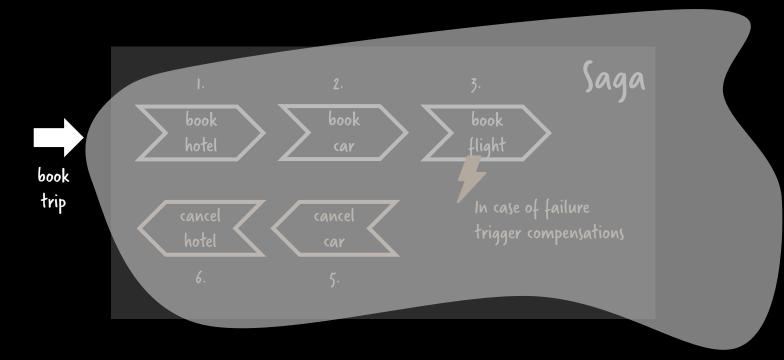
https://martinfowler.com/articles/201701-event-driven.html



The danger is that it's very easy to make nicely decoupled systems with event notification, without realizing that you're losing sight of that larger-scale flow, and thus set yourself up for trouble in future years.

https://martinfowler.com/articles/201701-event-driven.html

(lassical example





<u>Denis Rosa</u> Couchbase

If your transaction involves 2 to 4 steps, choreography might be a very good fit.

However, this approach can rapidly become confusing if you keep adding extra steps in your transaction as it is difficult to track which services listen to which events. Moreover, it also might add a cyclic dependency between services as they have to subscribe to one another's events.

https://blog.couchbase.com/saga-pattern-implement-business-transactions-using-microservices-part/

Microservice pioneers have become aware

Traditionally, some of these processes had been orchestrated in an ad-hoc manner using a combination of pub/sub, making direct REST calls, and using a database to manage the state. However, as the number of microservices grow and the complexity of the processes increases, getting visibility into these distributed workflows becomes difficult without a central orchestrator.



Netflix Technology Blog <u>Follow</u> Learn more about how Netflix designs, builds, and operates our systems and engineering organizations Dec 12, 2016 - 7 min read

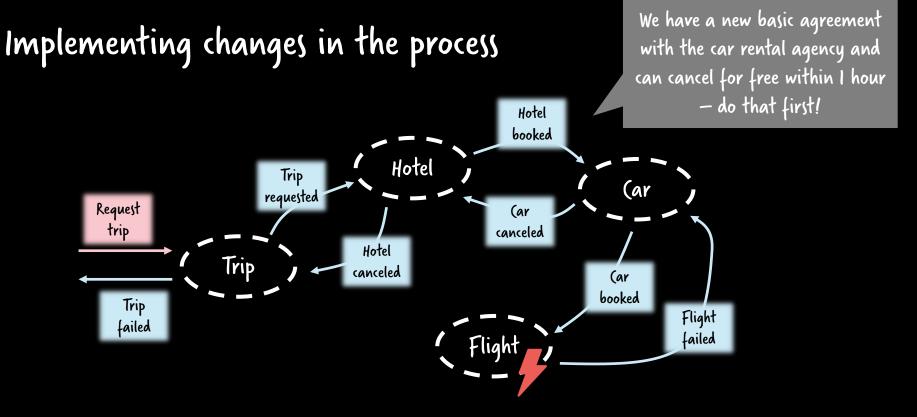
Netflix Conductor: A microservices orchestrator

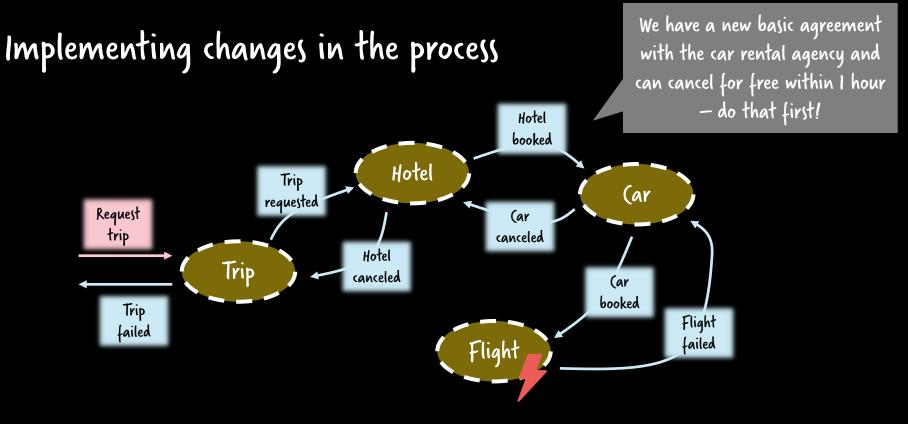
The Netflix Content Platform Engineering team runs a number of business processes which are driven by asynchronous orchestration of tasks executing on microservices. Some of these are long running processes spanning several days. These processes play a critical role in getting titles ready for streaming to our viewers across the globe.

A few examples of these processes are:

- Studio partner integration for content ingestion
- IMF based content ingestion from our partners
- Process of setting up new titles within Netflix
- · Content ingestion, encoding, and deployment to CDN

Traditionally, some of these processes had been orchestrated in an ad-hoc manner using a combination of pub/sub, making direct REST calls, and using a database to manage the state. However, as the number of microservices grow and the complexity of the processes increases, getting visibility into these distributed workflows becomes difficult without a central orchestrator.





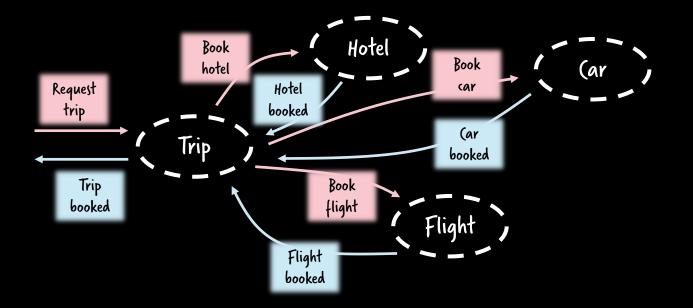
You have to adjust all services and redeploy at the same time!





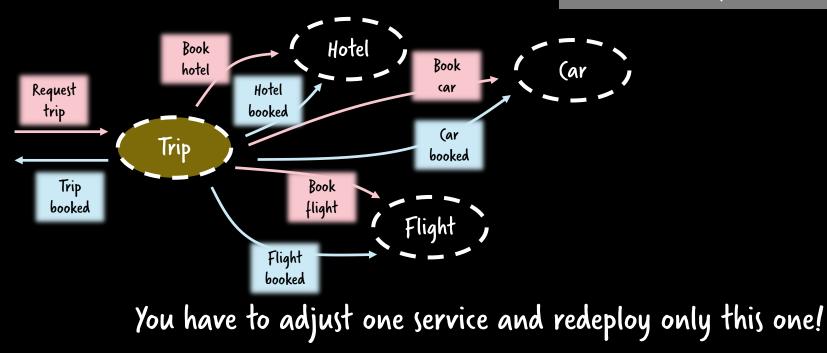
Photo by Lijian Zhang, available under Creative Commons SA 2.0 License and Pedobear19 / CC BY-SA 4.0

orchestration



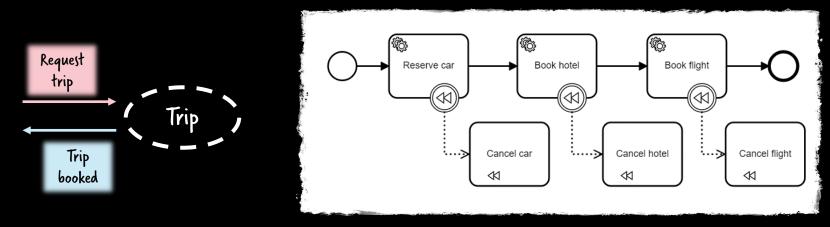
orchestration

We have a new basic agreement with the car rental agency and can cancel for free within I hour – do that first!

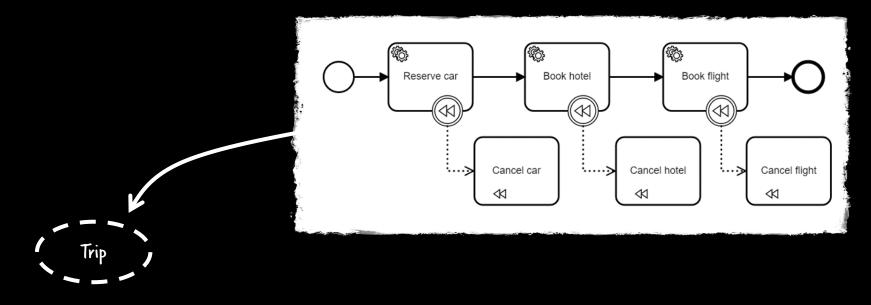


Describe orchestration with BPMN

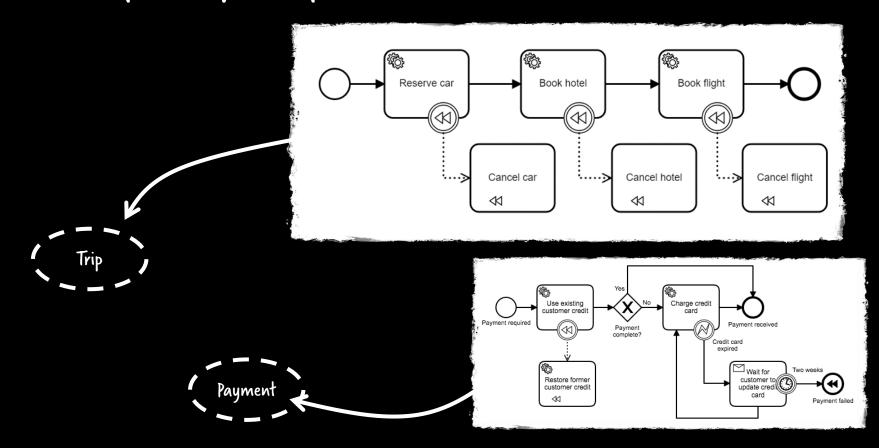


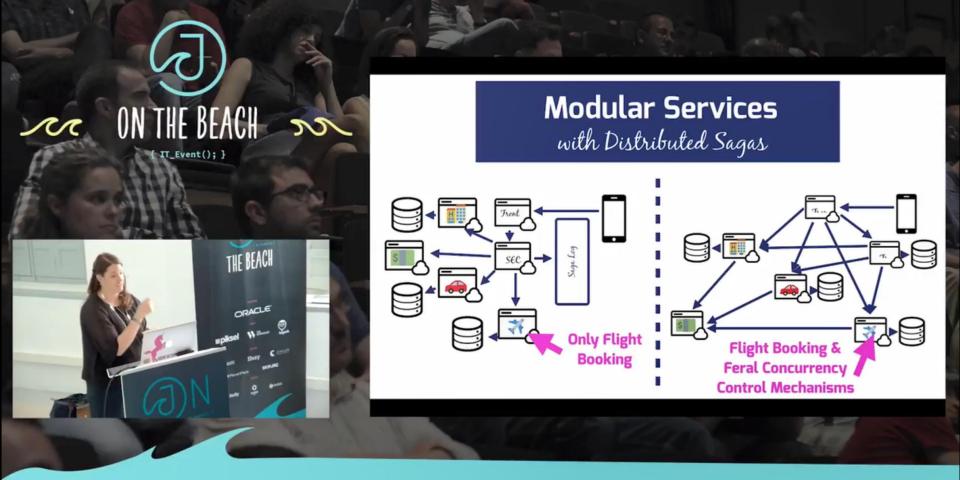


The workflow is part of the service



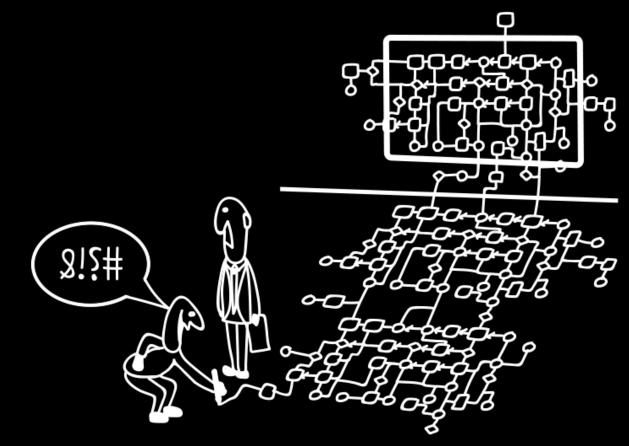
The workflow is part of the service





Caitie McCaffrey | @caitie

Graphical models?





<u>Clemens Vasters</u> Architect at Microsoft

http://vasters.com/archive/Sagas.html

Sagas 01 mode a

Clemenc Vactors

Rind

E Annie 1

Carrie .

Charles in

O Citva

Pressen

8 M.C.A.

O Role

10 10-10

d Declaration © Internal

BINC

"Boday Nac Boon a Hody day in come gars of the Twittenance debating the Saga gattern. Ac's transic, there are a few frameworker NTT out there that cos the term "Saga" for come framework implementation of a state reaching or workflow. The oble is, that's not what a Saga is a Saga is a fallow enangement pattern.

Sapaskased, box aka of the sail that in that gunrificative from-the dor search and configurably one-local to dos dataskased, box aka of databatic the nord transcelerations are done location and/or most beamders can't saily be handle sidely the databatic the nodel with 3-th ace server and halfing location to duration of the work's meased, a Saga agther work/inco indultation accession and halfing locations, so services after work's beamders.



The ploture shows a single Saga. If you books travel there any your ware a car and a horoland a flight. If you can't get all of them, if typebably narr works going, if takes any carrains that you can't entil a li of these providences a distribute a Soft traveaction, incosed you? This wan a soft for leading retrain any that is now both have to get form a recent atom and a loo have to cancel the and and for a horoland and for flight.

The a childrane or grouped is a sequenci bit housing digit that channels along the a calling due is type war, you can digit competent in each digit data can be that you can along be understand and an endpointee big the intension relation in white the campairing department and the campitation to the campit duping this thereins an white the campairing department and the scaled of g along that thereins and thy fails, it datas to g bad your then an each to each digit bad along that there is the data campined a addity that, it datas to g bad your then an each to each digit bad along that there is the campined a bad type and the many scale to each digit bad along the there is the term of the datas to bad you along the term of the many scale to each digit bad along the term of the campined a bad type and the many scale to each digit bad along the term of the there each digits along the term of te

If you've a be familiar with result you'l also notice that the anguable the coged synkic Reserving as restrained a large counced if you're beith values, because the restrainer and angung an more anne car service that there is high desmark Reserving a healt is diply more rich guot can anomenly backsoor of a recentation without genality until be before the casy lithtere often annex with a refund restration, as good waters to destate it.com

I preamed a Gier an Ghinde than you can run a calamedo application. It liberrarecthic model in code. Mind that it is a modug and not a framework i wrene this in loss than 60 minutes, so don't expect to reuse this.

The web programmetrue para exemploy rescring all gifts the descensive from an HBI and easeses three samplenes (independent "groesses" (antiby heard) that are a samir responde for hearding a gentrator field when The "groesses" (heard parameter and and head of heard heard) addresser for framma groegrees vorticing an and componentian work. The network weakhed in a devine all groefs.

in statis katistyteni() pressusy 20 20 statis odd Bob(string[) args) 20 (20 string statistics - new basistication interface)

> 7 no. Sandhannadartisti sitya (no. Sandhaninguanis), no. Sandhannan attaciasi sitya (no. Sandhaninguanis), no. Sandhannan attagistatu sitya (no. Sandhaninguanis) 20

// Degine these being sequences upper process with quark letters then processes - no. Intribution(ment); f model(a)(procedure-internation(a)(profile-int), model(a)(procedure-internation(a)(profile-int), model(a)(procedure-internation(a)(profile-int)); model(a)(procedure-internation(a)(profile-int)); model(a)(procedure-internation(a)(profile-int)); model(a)(profile-int); model(a)(profi

// hand aff its the first address Send(makingElip);

ninis and Box(Di ort, MadagElp radingElp) (// Wis is effectively the return disade

ferant (ur prans ir pransa) () () (mans inspirange(ri, radiogEig)

The activities each implement a reconstion step and an understep. Here's the one for care

allow hear-alleriations - installing
 (a)
 (b)
 (a)
 (b)
 (b)
 (b)
 (b)
 (c)
 <li(c)
 (c)
 <li(c)
 <li(c)
 <

pálla samia linkag bilek(linkõis antõis)

 Details initiation ("hearing as"); are are - untilent inputting (untility); are resentation - real last (untility); bendia initiation ("hearing are (it)," resentation(); bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation();
 bendia initiation("hearing are (it)," resentation("hearing are (it)," resentation("hearin

where we limit applied, we limitedly () reservabledly, reservabledly (χ_{0}

Constantivities ("Constitution" (8)", reservation(8) return inter **(**),

Clemenc Vactorio

R Ind

D Tanke D Tanke

Continue 1

BINC

Distant October

d Darks of an

B 16046

O Role

The adapting large-restrict/means the restring dip. The restring dip Windbabbs/ (Virtual), person that it is year of the restrict of the restr

ph (when we bely have been been as the

St. Kim hallingEig realies), Bachtlerkage angleteilierkage - nu Bachtlerkage(), realies), SacriterkSeer realierkSee - nu SacriterkSee(), AND MADE DIS salis hadaptic (Inventionistics and as) ferent (un untilles in untilles) No. restartilizes Insura (artifica) participant Scheduler pri (miure tria continet Con Cont=
by) philis had biblingens ph (when this ampleteriaritage face + 8;) while and Presselent of these bookstones You as includent for the later or sevenible - this settimber $\log n_{0}()$ or attribut - (attribute $\log n_{0}()$ an real in addition (according) of creatile in with nia amfahalimtaga Padomatik) and therefore a Secola Science (Secolar (S)), a Percepty network Address nitis ini Propresi in restartCare Parts (John Ry (pr) of COLORINGTON (Setting) (Address Presidenters (Mission) Address Park (Setting) path and path at of Chick Schwarzen these are involving and a function (a) ar sararitha - this anglaballaritags.he()) ar aslaidy - (laticity)laticate: Destellation un Inter internet Den Jahl die Den in

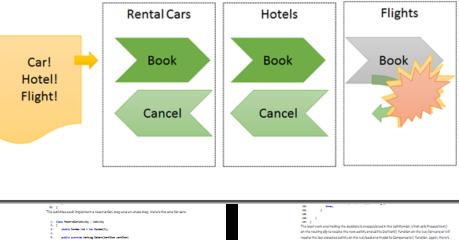
this and a sector



<u>Clemens Vasters</u> Architect at Microsoft

http://vasters.com/archive/Sagas.html





Canada IndiaLine ("Reaming car");

million lines

ur ur - urbin.igunig("ubilitys")) ur restuliete - reliet(2000)) trade bilitiet("teanet ar (0)", restuliet())

um mamatianti - Dan bash("mamatianti"); Genda Sebelan("Genditai an (8)", mamatianti);

return neutienticationis, neutienticatio 6 6 Insurrunticatio, recencidade 3 Vic-

public committe hard Compressive(Service Den, NationpEle realizeDia)

The lab have an array of the academic array calculate in the administration of the lab Production of the lab P



<u>Clemens Vasters</u> Architect at Microsoft

http://vasters.com/archive/Sagas.html

Sagas 02 rooks met

Clemenc Vactors

B Ind

E Annie 1

Carrie .

Charles in

C Internet

Ocina

Designation

BINC

Today has been a fiely day in some parts of the Twittenerse debating the Saga gattern. So it stands, there are a few frameworkfar. NET out there that each the term "Saga" for some framework implementation of a state machine or workflow. Theobie is, that there what a Saga is a Saga is a fallow enaugement pattern.

Sagac area our of the next lists in that garried with (sing-the dir re-next direc (angle hall) years), part holds databased, but a last of admictable management loant on angler must be under the all years handle ut degrin the datababb model with 3-bit and and main and halfing load for the datababb model work thereas, a Saga agther work hand half and management of the card of a the hard model and the performant and a common datababb model.



The gleane draws a drags 2 age, it you beaks most life samp your wares a arrow is here it and a fight. If you can't ge rail of them, it's probably more the physic Vision using samp in that you can't enders if at these generatories and a dimbane do the manadam. It make is, out That as an a sith fight to be doing reservaand that insure both how to genterm a neurration and also how to an and the and on for a here land and for this can be done the same second on a site is not a same the - and one for a here land

The a childrane grouped is a sequenci jab locating digit hardshareda along the acatily due is Type war, go can digitumg the societ digit beams that they can only be unkerstead and an endpoints by the knowled relation. This was acably anglesics, baseds a ream's diff a anglesites to the acatig due lange with Mineration analysis are applied as the acating digital distribution. The acating digital genesis the Mineration analysis are applied as the acating digital distribution to the acating Within an acatity to fit, it dances to poly years then an action excited gita distributions the later amplied as ability amplied and the acating distribution actions.

If you've a be familiar with result you'l also notice that the anguable the coged synkic Reserving as restrained a large counced if you're beith values, because the restrainer and angung an more anne car service that there is high desmark Reserving a healt is diply more rich guot can anomenly backsoor of a recentation without genality until be before the casy lithtere often annex with a refund restration, as good waters to destate it.com

I preased a Gier an Gehole that you can run a calancelo application. It illuctratective model in order. Mind that it is a modup and not a framework i wrene this in loss than 60 minutes, so don't expect to rease this.

The well-programmetric part energies/instructing dig (2) the discontrast in the energies of th

1: static bibliches] presson 2: 5: 6: (bibliches] presson 6: (bibliches] presson 6: (bibliches] 6: (circles) presson 6: (circ

 Informationalisticity(n) information(), no informationalisticity(n) information() 20

// Longine these body explanation processes with quark letters the processes - no. intribution/proced f no. intribution/procedment and anticipation(grad), no. intribution/procedment and anticipation(grad), no. intribution/procedment and anticipation(grad).

// hand off in the first address Beni/makingElip.frequencial, makingElip);

nite and first(int only having file reading file)

ferant (ur press in presse) H (press inspiring)(ut, redirgEig)

The activities each implement a reconstion step and an understep. Here's the one for care

state hear-all-faithing - lating
 (
 shift hear minimum in the hear (2);

palls service listing beint (interview until etc.)

Desubs.initiative("hearsing as");
 ser ser - serbites.inguesis("setbit(get");
 ser resentationis - resides(38888);
 Desubs.initiative("hearsist ser (0); resentation(b);

where we likely (i.i., we likely such (() was also field) was added by χ_{0}

public summits had Dependentianing Day, havingDip readingDip (ar resenation() - Day Ana/Affreenation())

Denails initiation ("Constitution" (0)", reservationity return integ 20 (pr (return no. 24("shot" doublestication");) 26 ; ; 26 ; ;

Clemenc Vactor

R Inc

0.54

COMP.

Distant October

d Darks of an

Bischer Bischer iii - 3 The ada high apport adaly the range has morting die. The scantup die is "savitable" (if views, present that is its add in the adaly gives of the range has a first order and the adapt of the savitable. These there is a first of the adapt gives of the range of the savitable of the savitable of the proceeding die forwards, it moused or balances of rankets. The scantup die is a first order die has a range and for balances the mouse is readered and the savitable of the scantup die is also as a first first dimensioner market has a savit first and address and for the adapt die water and the savitable of the savitable of the savitable of the scantup die is the scantup die water and the savitable of the savit

vingCia) Bashtietlage angletalietlage - na Bashtietlage[]; Quartietlage multietlage - na Quartietlage[]; satis hadapEis(ShownaterianOres antOres) ferent (un untilles in untilles) No. restartilizes Insura (artifica) participant Discription pri (miure tria continet Con Cont=
by) philis had biblingens ph (when this ampleteriaritage face + 8;) while and Presselent of these bookstones You as building for the later an real in addition (according) of creatile in with nia amfahalimtaga Padomatik) and therefore a Secola initialize ("Inspire" (0)", a Percepty network Address patter tet Program in restartCare Parts (John Ry (pr) of COOPeration (Setting) (Address Presidenters (Mission) Address Park (Setting) path and path at of Chick Schwarzen these are includent information (); Our - this ampletelleritage Party (- anticity - (latticity)latticities.dre and and neuro anticity Companyis(surrentDan, Mile); on therein d Constantivity of Taxable (UT), a Presentaalchrosofter() The local work and mailing the deddone leancageular on the routing dig to receive the next activity and call h e wayforward or will recolve the last executed activity on the way backand in function dgain, there's nothing contraited here; all that workhinger on the n ethilties and their everation is completely eligent Annual Case Section Interiori, NatiogElips and

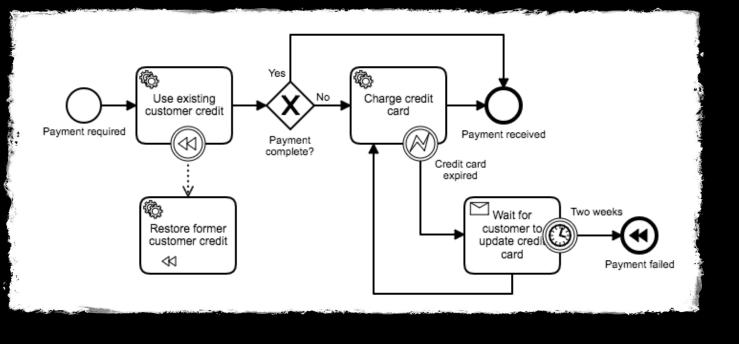
BPMN

Business Process Model and Notation

ISO Standard



Living documentation for long-running behaviour

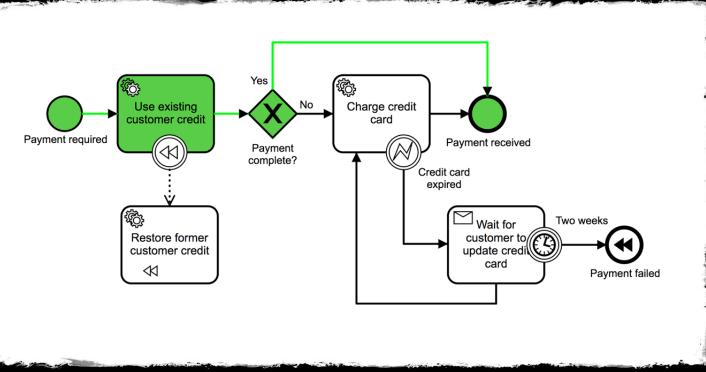




BY EXAMPLE

MANNESS

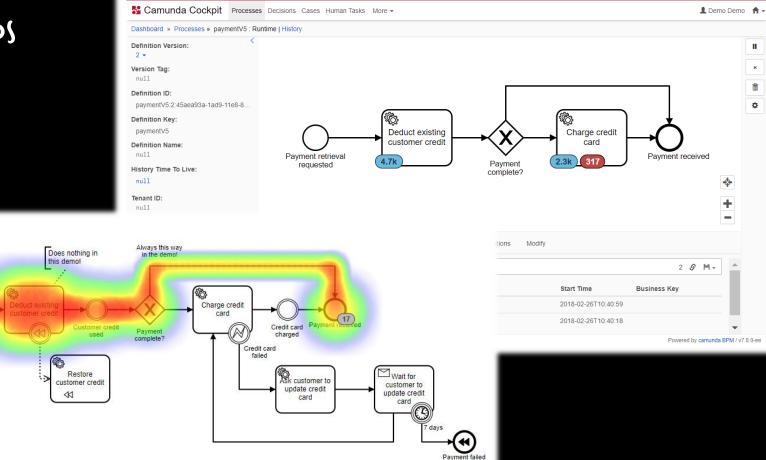
Visual HTML reports for test cases



BizDevops

Payment retrieval

requested



Ш

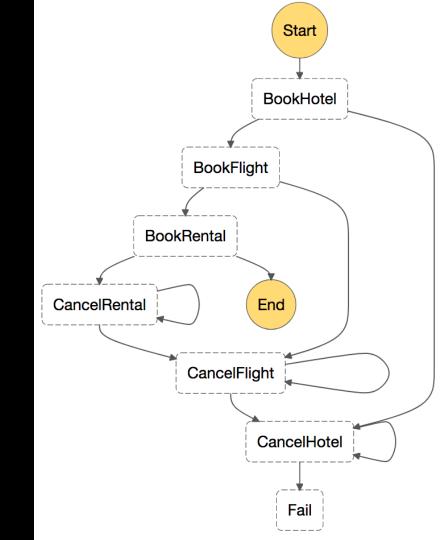
×

ŵ

¢

Saga with AWS Step Functions

https://theburningmonk.com/2017/07/applying-thesaga-pattern-with-aws-lambda-and-step-functions/



Thoughts on the state machine | workflow engine market



Thoughts on the state machine | workflow engine market

Stack Vendors, Pure Play BPMS Low (ode Platforms

PEGA, IBM, SAG, ...

(amunda, zeebe, jBPM, Activiti, Mistral, ...

> oss Workflow or orchestration Engines

Integration Frameworks

Apache (amel, Balerina, ...

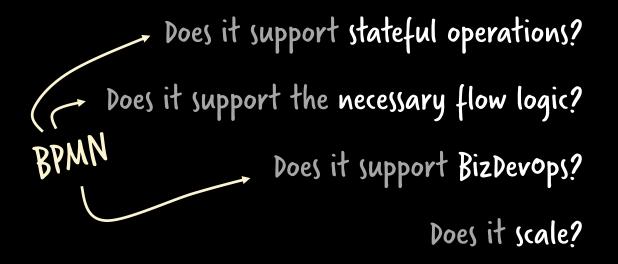
Homegrown frameworks to scratch an itch

Vber, Netflix, AirBnb, ING, ...

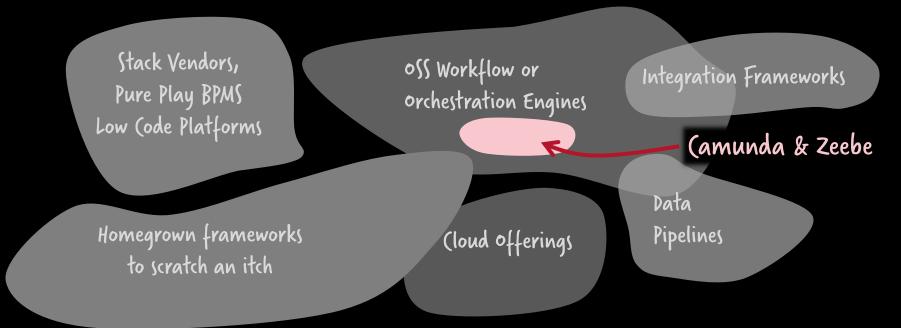
(loud offerings

AWS Step Functions, Azure Durable Functions, ... Data Pipelines

Apache Airflow, Spring Data Flow, ...



My personal pro-tip for a shortlist ;-)





- Grown ups don't use distributed transactions but eventual consistency
- Idempotency is super important in distributed systems
- Some consistency challenges require state
- Know some strategies
 - Stateful retry & cleanup
 - Saga / (ompensation
 - Apologies



- Contact: <u>mail@berndruecker.io</u> @berndruecker
 - Slides: <u>https://berndruecker.io</u>
 - Blog: <u>https://medium.com/berndruecker</u>
 - Code: <u>https://github.com/berndruecker</u>



https://www.infoworld.com/article/3254777/ application-development/ 3-common-pitfalls-of-microservicesintegrationand-how-to-avoid-them.html

Info

https://www.infoq.com/articles/eventsworkflow-automation



https://thenewstack.io/5-workflow-automationuse-cases-you-might-not-have-considered/

