VoltDB
Things you learn as you massively scale…

David Rolfe
Director of Solution Architecture, EMEA

Tom Howcroft
Director of Sales, EMEA

17-Jul-18

VOLTDB

Scaling at the Architectural Level…

How many servers will you need to start?

- HA implies more than one machine
  - With only 2 nodes you need 100% spare capacity
  - With 3 50% spare, with 4 33% spare…
  - So: Don’t assume a cluster of two ‘monster’ servers is optimal.

- Something will be a driving factor. Do not guess this – measure it!
  - HA
  - RAM
  - CPU
  - Network

- You may not be able to dictate the size of servers…
  - Example: AWS may require a certain size node for an adequate network
  - Reality check: ”Someone Else’s Cloud” will have its own selection of available size.

How many servers will you need eventually?

How many spare copies do you need?

- As the number of machines goes up the chance of a failure goes up…
- You have 1 spare copy of data but what if both copies are lost because you lost 2 out of 20 servers?

- Eventually you'll need two spares. When is dependent on your level of paranoia…
- Hybrid approach is to have 'wallflower' nodes that will rapidly join cluster
  - Reduces time spent with only 1 copy from hours to minutes

- Do you reject peak traffic or size for it?
  - You do have a plan for peaks, don't you?

Will you need multiple sites?

- Historically Active-Active was ‘science fiction’

- Now it’s a common requirement

- Motivation
  - Survivability
  - Latency
  - Ego

- Doesn’t help you scale
  - Everybody has to find out about every transaction everywhere
  - Going from Active-Active to Active-Active-Active implies extra work even if new site does nothing

How do you partition the data?

You mean we have to partition?

- For low latency environments with writes partitioning is unavoidable.
- Pick the least awful partition key…
- VoltDB’s Materialized views can help…
- Eventual Consistency isn’t
  - Side effects of inconsistent reads will propagate way beyond the database before data is made consistent.
- Do you reject peaks or size for them?

Broader Implications…

- System is too complicated to do testing on a laptop:
  - RAM
  - Network
  - CPU
  - …all non trivial
- Development and Testing costs will spike
- Problems with behavior changing between Dev and Test
- Problems with emulating connected systems in Test

VOLTDB

Scaling Write Intensive Workloads…

Scaling “Writes” isn’t like scaling “Reads”…

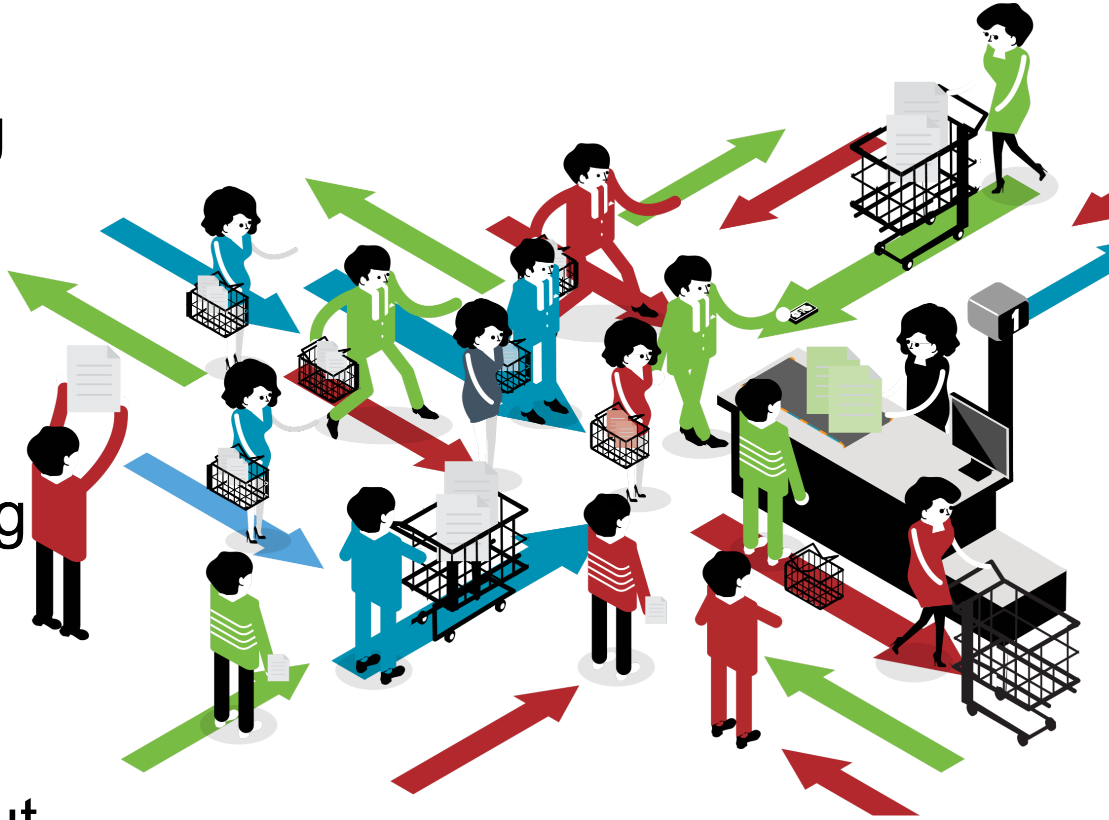
- Traditionally we scale by adding more of *whatever is most needed*.
- So commodity hardware is great at scaling reads, as reads need CPU, RAM etc
- Some writes scale well – e.g. if they are inherently unique and disconnected from anything else.
- But if writes need to be ACID we can’t simply have two separate updates to two copies in two places.
- The bottleneck is not a physical resource.
- In this case ”*Whatever is most needed*” is the data itself.
  - Implies you can’t solve this problem with hardware

If we tried DB write strategies in a supermarket…

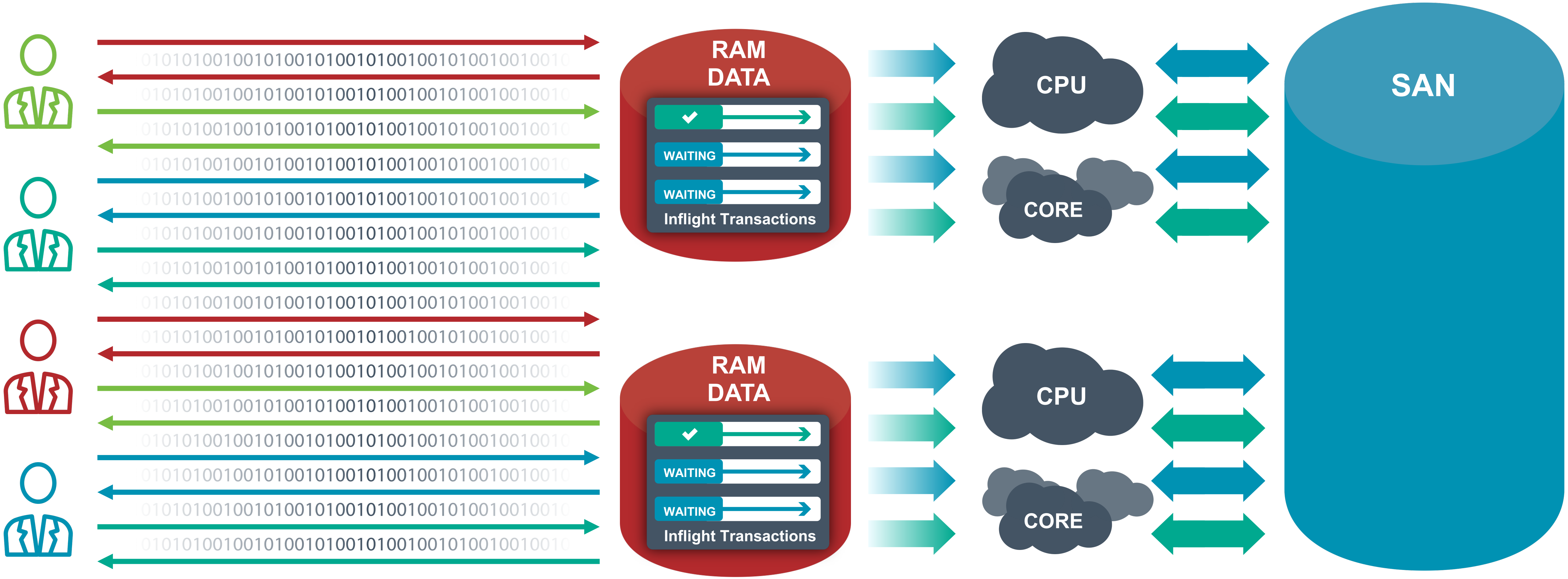
Row Level Locking: Nobody can touch the Orange Juice shelf or any other shelf I'm taking things from until I've finished shopping and checked out!

Eventual Consistency: I take Orange Juice, then pay for it, but it vanishes from my shopping cart and moves to someone else's as I put my bags in my car. The staff deny this happened.

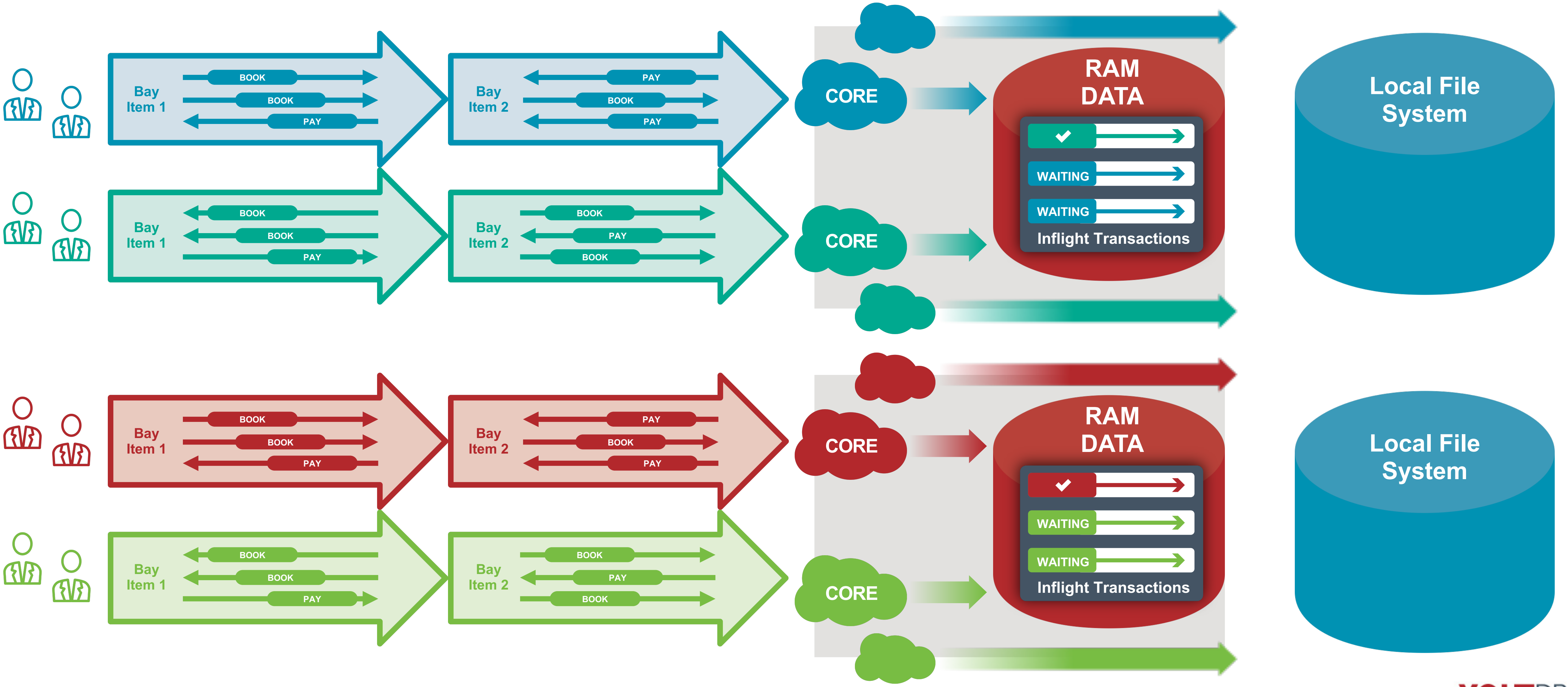
Optimistic Updates: I buy my Orange Juice but are pulled over by security as I attempt to drive away. They refund my money and take the Juice off me, then tell me to try again.



RDBMS - What Actually Happens – Part 2



How VoltDB works



VOLTDB

Scaling in the real world…

Or “6 things I wish I knew before I started”

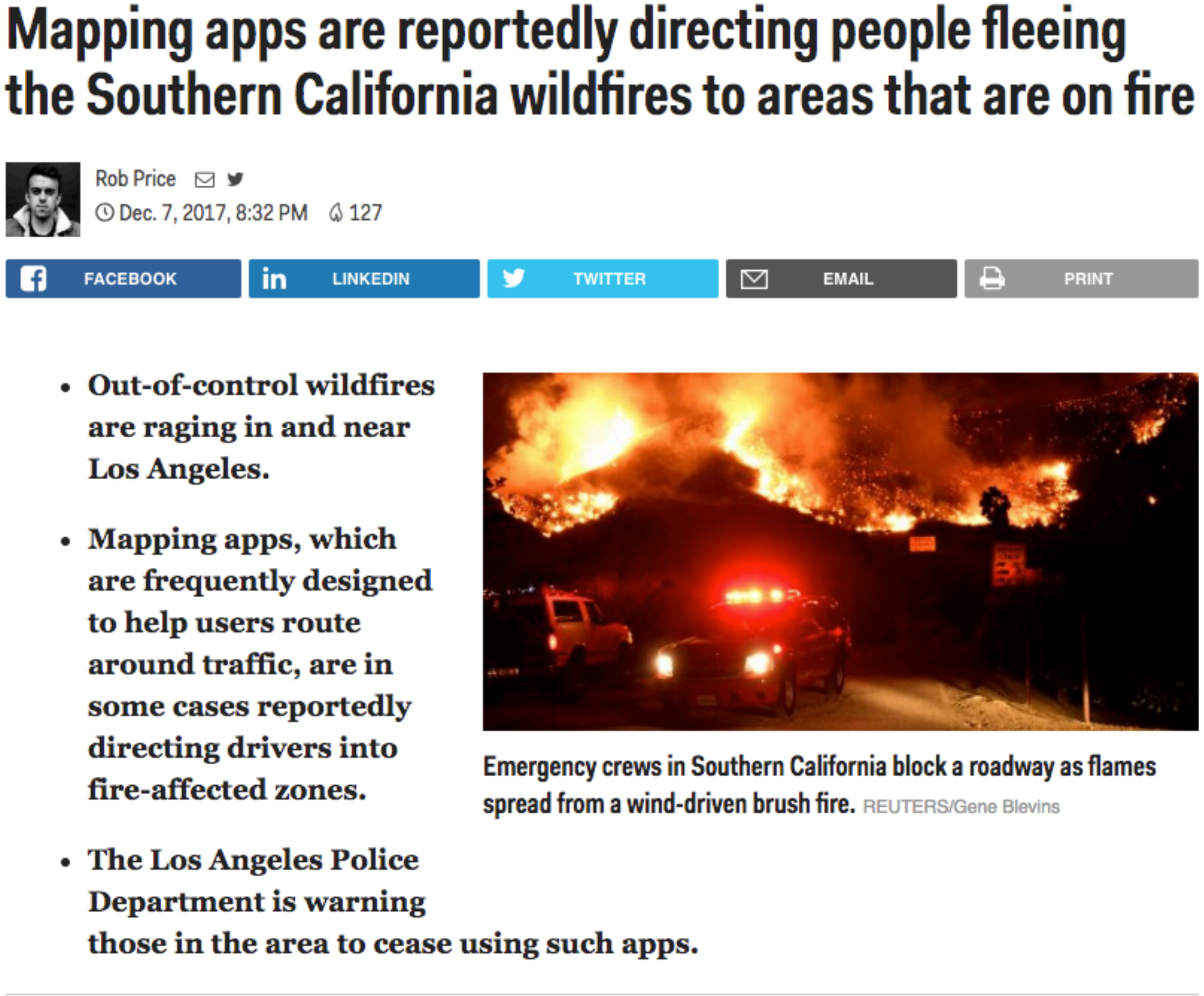
1. Ludic Fallacy

“Ludic Fallacy” – Mistaking a game for reality…

Our model can never perfectly match reality.

Which means that no matter how ‘well trained’ it is, there will be a scenario which the model oversimplifies or otherwise fails to cope with.

1. Ludic Fallacy – An Example



Mapping apps are reportedly directing people fleeing the Southern California wildfires to areas that are on fire

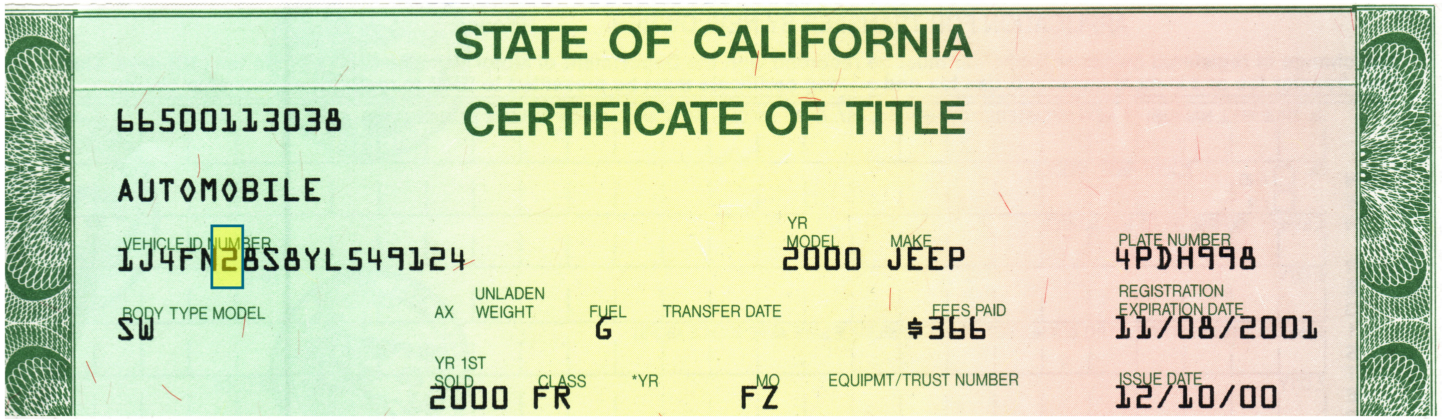
Rob Price

Dec. 7, 2017, 8:32 PM 127

- Out-of-control wildfires are raging in and near Los Angeles.
- Mapping apps, which are frequently designed to help users route around traffic, are in some cases reportedly directing drivers into fire-affected zones.
- The Los Angeles Police Department is warning those in the area to cease using such apps.

Emergency crews in Southern California block a roadway as flames spread from a wind-driven brush fire. REUTERS/Gene Blevins

2. Your Data Is Always Slightly Wrong



STATE OF CALIFORNIA

66500113038

CERTIFICATE OF TITLE

AUTOMOBILE

VEHICLE ID NUMBER 1J4FN2BS8YL549124

YR MODEL 2000

MAKE JEEP

PLATE NUMBER 4PDH998

BODY TYPE MODEL SW

AX

UNLADEN WEIGHT

FUEL G

TRANSFER DATE

FEES PAID $366

REGISTRATION EXPIRATION DATE 11/08/2001

YR 1ST SOLD 2000

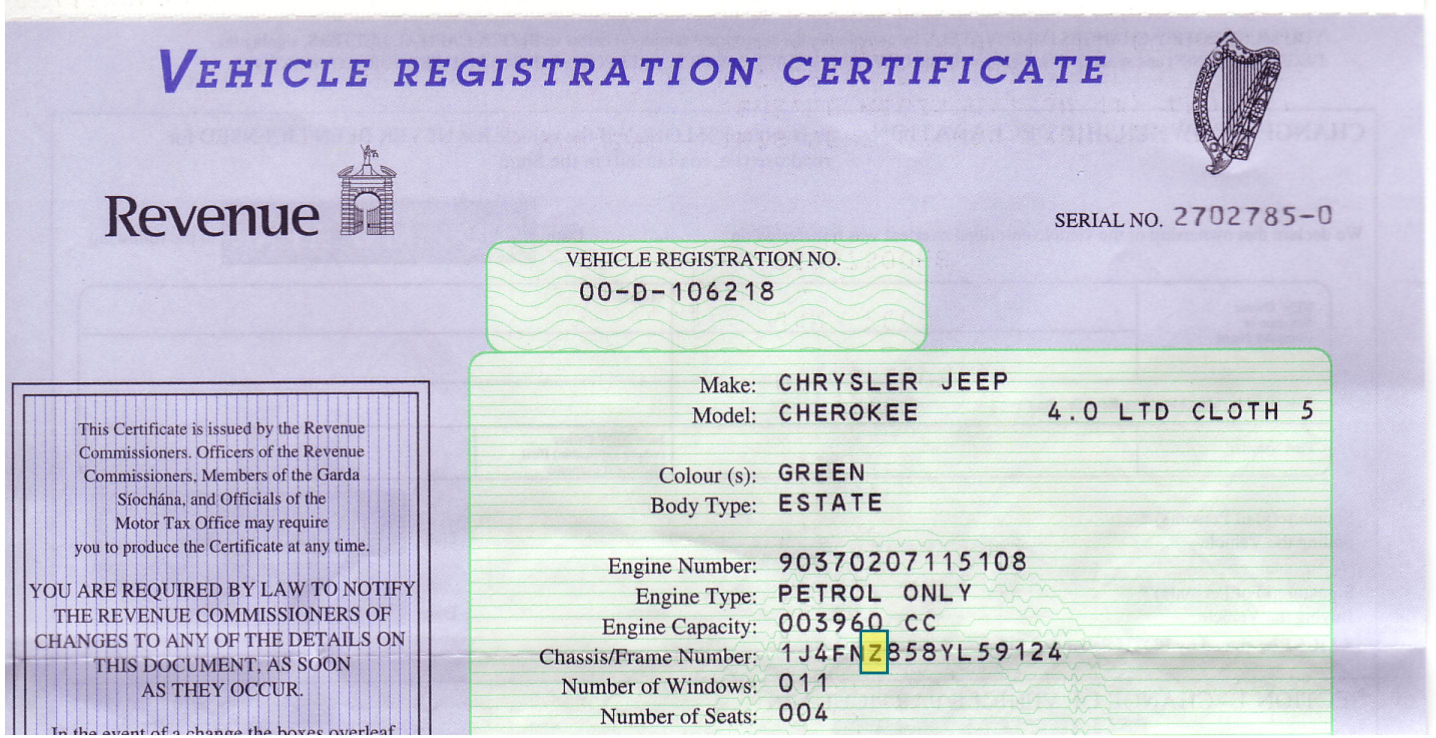
CLASS FR

*YR

MO FZ

EQUIPMT/TRUST NUMBER

ISSUE DATE 12/10/00



VEHICLE REGISTRATION CERTIFICATE

Revenue

SERIAL NO. 2702785-0

VEHICLE REGISTRATION NO.
00-D-106218

Make: CHRYSLER JEEP

Model: CHEROKEE 4.0 LTD CLOTH 5

Colour (s): GREEN

Body Type: ESTATE

Engine Number: 90370207115108

Engine Type: PETROL ONLY

Engine Capacity: 003960 CC

Chassis/Frame Number: 1J4FNZ858YL59124

Number of Windows: 011

Number of Seats: 004

This Certificate is issued by the Revenue Commissioners. Officers of the Revenue Commissioners, Members of the Garda Síochána, and Officials of the Motor Tax Office may require you to produce the Certificate at any time.

YOU ARE REQUIRED BY LAW TO NOTIFY THE REVENUE COMMISSIONERS OF CHANGES TO ANY OF THE DETAILS ON THIS DOCUMENT, AS SOON AS THEY OCCUR.

In the event of a change the boxes overleaf

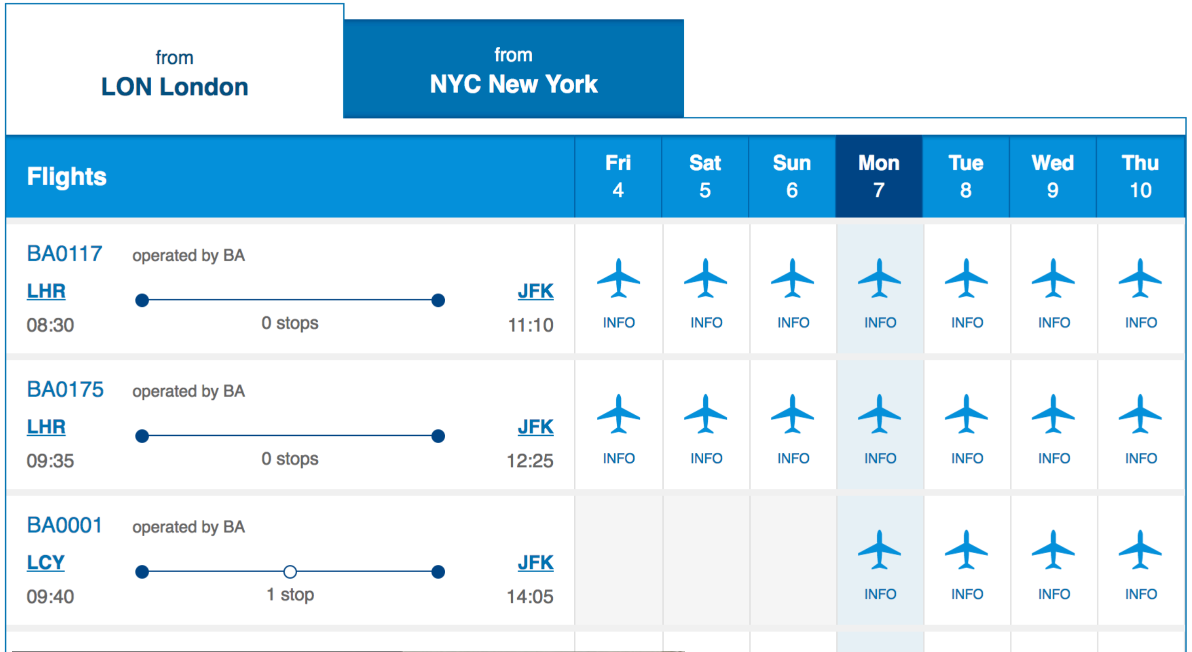
Real world data streams are always imperfect.

Example: The chassis / VIN number of an automobile can never change, ever!

Information about the 'ghost' vehicle went was sent to the police, insurance industry, stats agency….

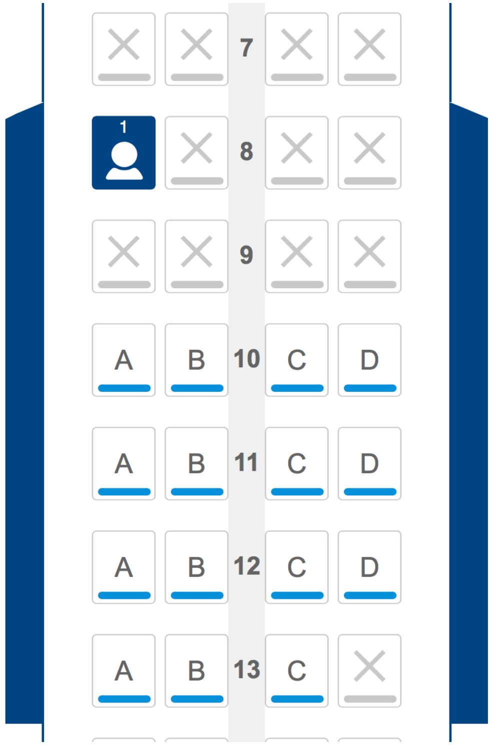
3. Merging multiple data streams is hard

Goal: Predict flight delays.

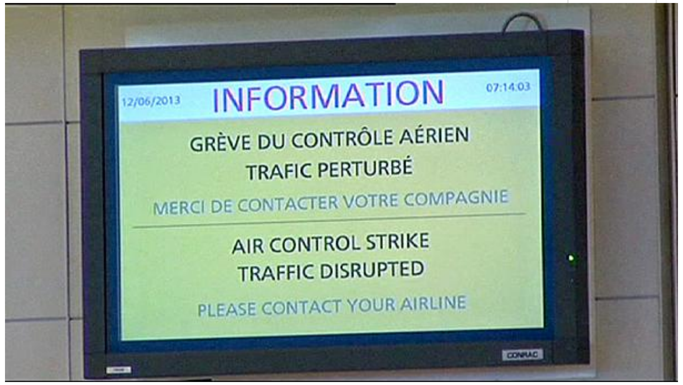


from **LON London** | from **NYC New York**

Flights	Fri 4	Sat 5	Sun 6	Mon 7	Tue 8	Wed 9	Thu 10
BA0117 operated by BA LHR 08:30 — 0 stops — JFK 11:10	INFO	INFO	INFO	INFO	INFO	INFO	INFO
BA0175 operated by BA LHR 09:35 — 0 stops — JFK 12:25	INFO	INFO	INFO	INFO	INFO	INFO	INFO
BA0001 operated by BA LCY 09:40 — 1 stop — JFK 14:05				INFO	INFO	INFO	INFO



✗	✗	7	✗	✗
1	✗	8	✗	✗
✗	✗	9	✗	✗
A	B	10	C	D
A	B	11	C	D
A	B	12	C	D
A	B	13	C	✗



12/06/2013 INFORMATION 07:14:03

GRÈVE DU CONTRÔLE AÉRIEN
TRAFIC PERTURBÉ
MERCI DE CONTACTER VOTRE COMPAGNIE

AIR CONTROL STRIKE
TRAFFIC DISRUPTED
PLEASE CONTACT YOUR AIRLINE

"The Late Arrival Of The Incoming Aircraft"

Raw TAF

```
KJFK 070809Z 0708/0812 36004KT P6SM SCT025 BKN040
  FM071400 04009KT P6SM SCT035 BKN050
  FM071800 15010G15KT P6SM SCT035 BKN050
  FM080100 09009KT P6SM SCT030 BKN100
  FM080900 05005KT P6SM SCT020 SCT100
```

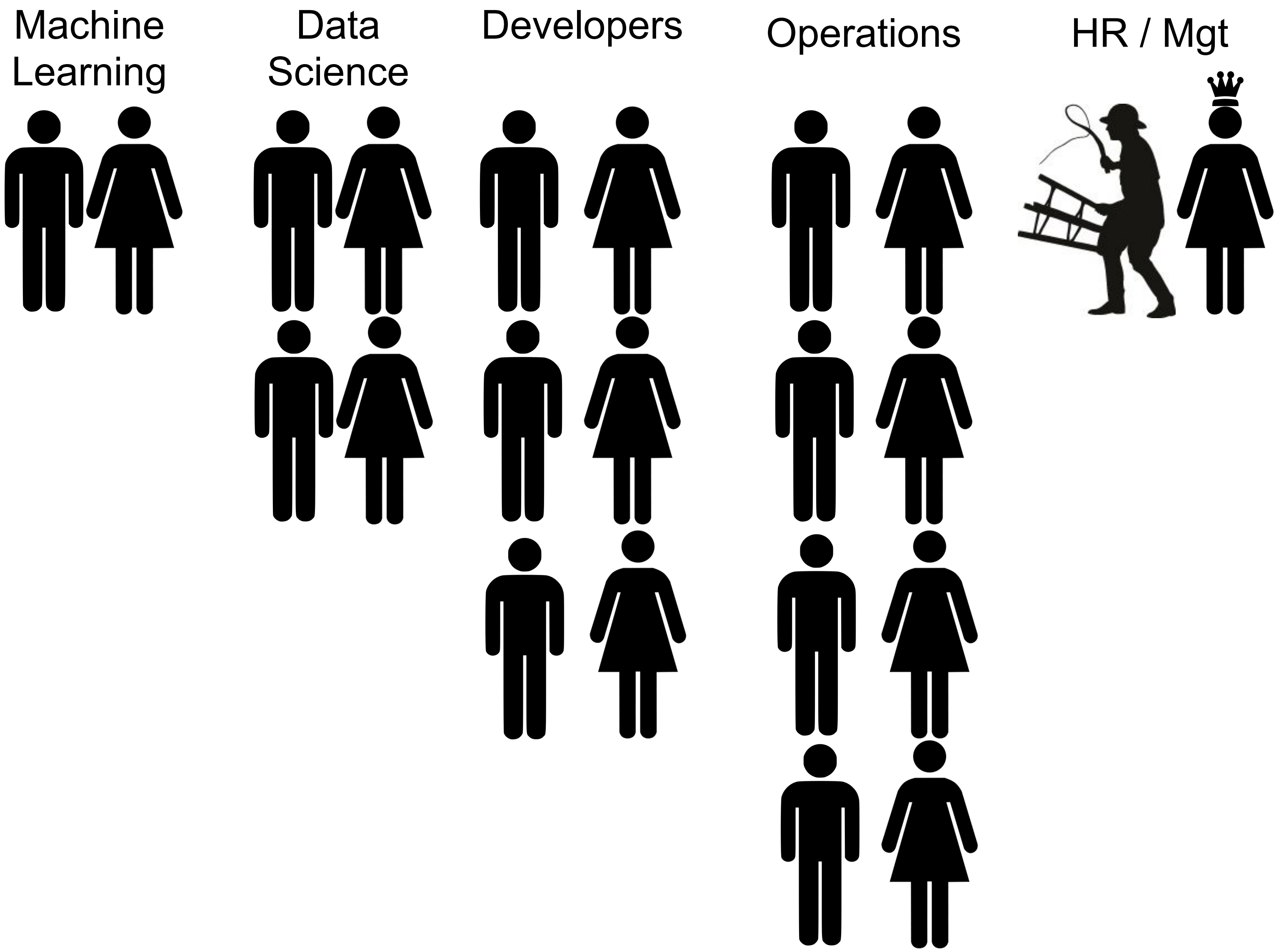
Raw METAR

```
KJFK 070951Z 35006KT 10SM FEW060 BKN250 13/11 A3000 RMK AO2 SLP159 T01280106
KJFK 070851Z 35005KT 10SM FEW060 BKN250 12/11 A2998 RMK AO2 SLP152 T01220106 53013
KJFK 070751Z 36004KT 10SM FEW055 BKN250 13/11 A2996 RMK AO2 SLP146 T01330106
KJFK 070651Z 36008KT 10SM SCT024 BKN055 14/11 A2996 RMK AO2 SLP144 T01440111
```

VOLTDB

4. As volumes increase, life will get much harder.

5. Loading the data will never finish



6. What happens if time is of the essence?

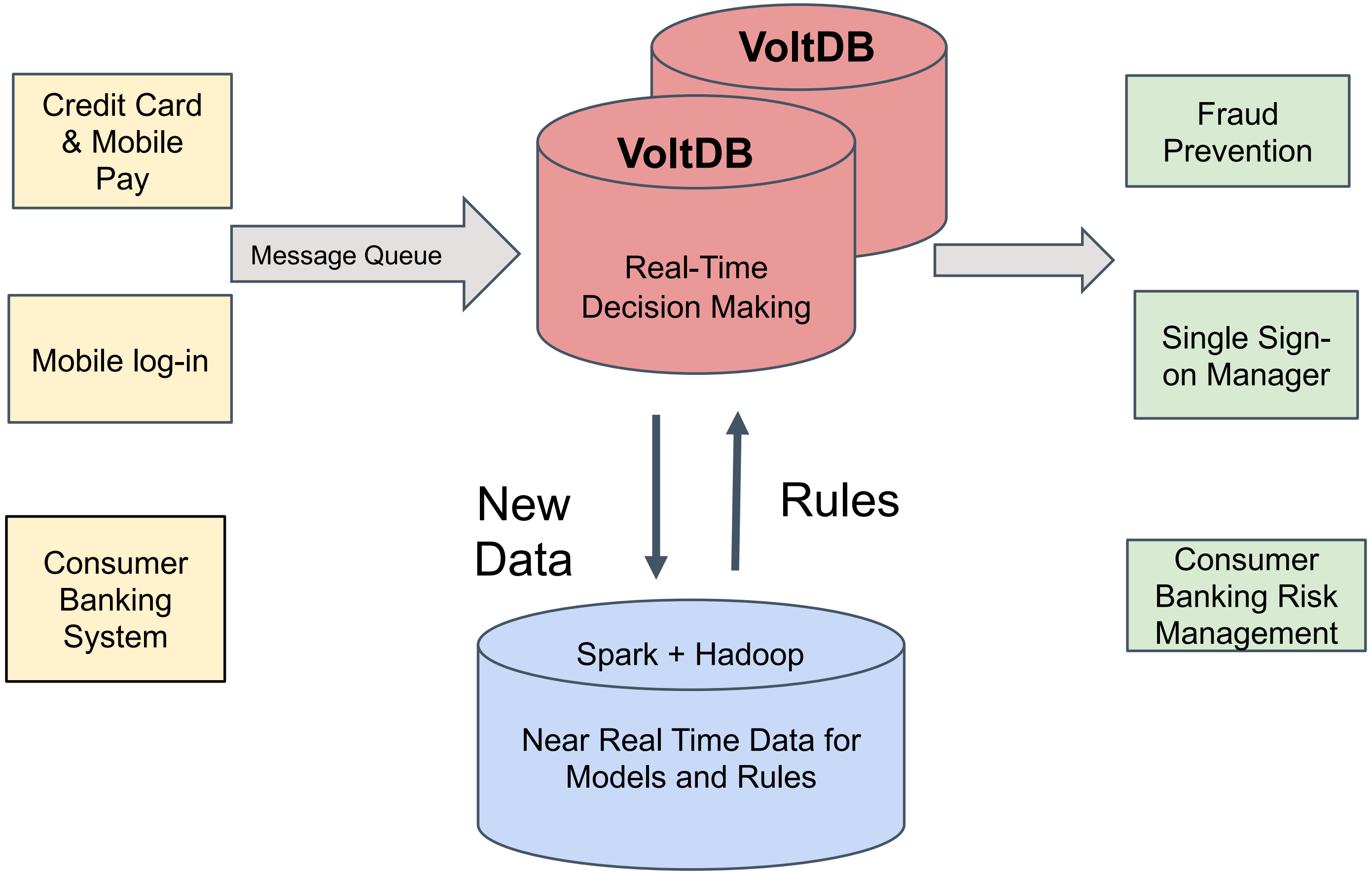
Traditional Batch / Hadoop Speed: 30 Minutes

Web Server : 3-7 Seconds

Spark / Kafka : 1-2 Seconds

Traditional OLTP: 5-50 ms

5G Phone Network / VoltDB: 1ms



Application/Use Case

- Fraud Prevention
- Single sign-in of all Huawei phones
- Consumer banking risk management

Why VoltDB?

- > 50% reduction in fraud cases
- > $15M/year saved from fraud loss
- 10k complex Transactions Per Second
- 99.99% transactions finish < 50ms
- 10x better performance than traditional fraud detection