Big Data in Production: Lessons from Running in the Cloud

Marvin Theimer

Amazon Web Services

SPARK SUMMIT 2016

DATA SCIENCE AND ENGINEERING AT SCALE

JUNE 6-8, 2016 SAN FRANCISCO

From Prototype to Production

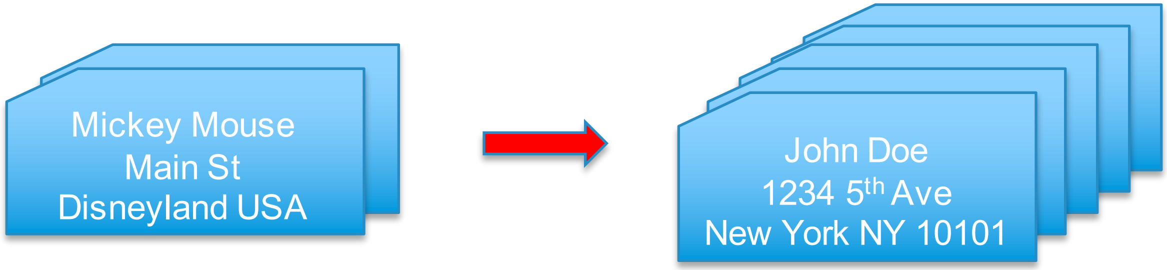
Wow!! I got it to work!

How soon can I have a million of those?

Production “ilities”

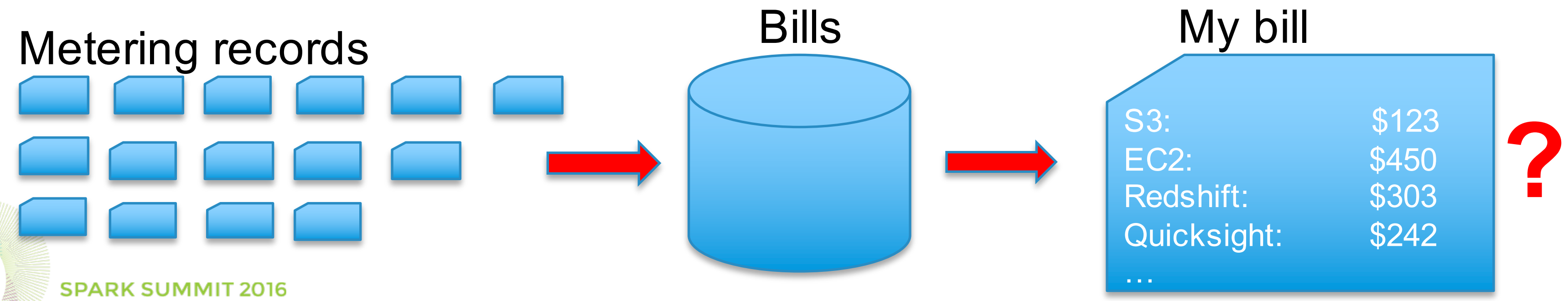
- **Scalability**: many jobs, running day-in and day-out
- **High availability**: no excuses for not delivering results
- **Maintainability**: seamless upgrades, security patches, backups, …
- **Evolvability**: the new person, a year from now, can add new features to your production system
- …

Test Data → Valuable Data



Mickey Mouse
Main St
Disneyland USA

John Doe
1234 5th Ave
New York NY 10101



Metering records

Bills

My bill

S3:	$123
EC2:	$450
Redshift:	$303
Quicksight:	$242
…	

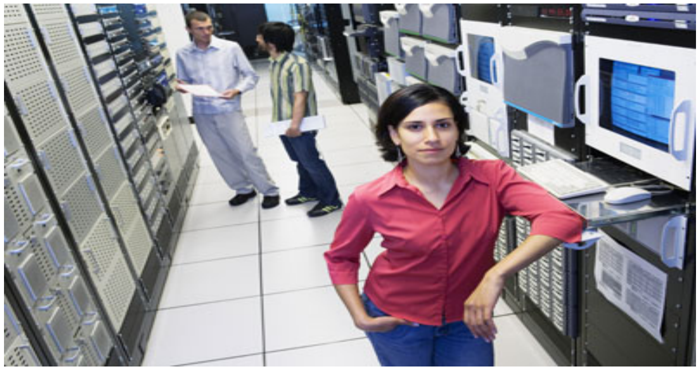
Algorithmic efficiency → "mundane" efficiency

Constants matter

vs.

The never-ending battle against waste

Fine-Grained Control vs. Ease-of-Use



vs.

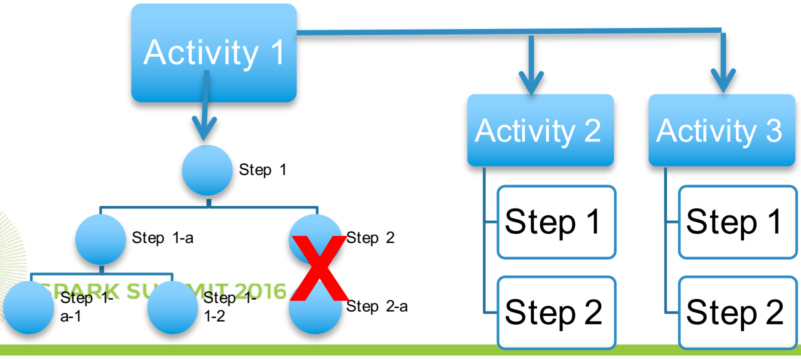
You’ll never believe the variety of workloads they’ll feed you

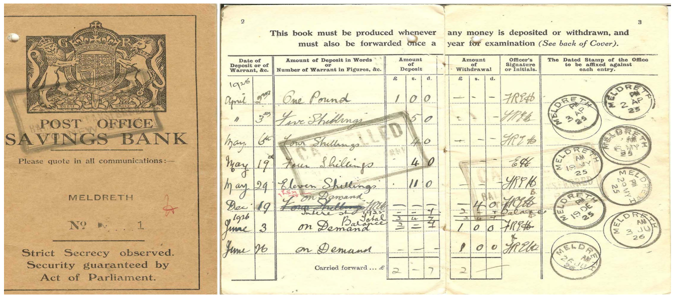
Scale → Automation

High Availability

Single points of failure vs. “large scale events”

Where did it go wrong, when, and why??





POST OFFICE
SAVINGS BANK

Please quote in all communications:—

MELDRETH

Nº [illegible] 1

Strict Secrecy observed.
Security guaranteed by
Act of Parliament.

2 | 3

This book must be produced whenever any money is deposited or withdrawn, and must also be forwarded once a year for examination *(See back of Cover)*.

Date of Deposit or of Warrant, &c.		Amount of Deposit in Words or Number of Warrant in Figures, &c.	Amount of Deposit £	s.	d.	Amount of Withdrawal £	s.	d.	Officer's Signature or Initials.	The Dated Stamp of the Office to be affixed against each entry.
1925										
April	2nd	One Pound	1	0	0	—	—	—		MELDRETH PM 2 AP 25
"	3rd	Five Shillings		5	0	—	—	—		MELDRETH PM 3 AP 25
May	6th	Four Shillings		4	0	—	—	—		MELDRETH AM 6 MY 25
May	19th	Four Shillings		4	0	—	—	—	E.H.	MELDRETH AM 19 MY 25
May	29	Eleven Shillings		11	0	—	—	—		MELDRETH PM 29 MY 25
Dec	19	on Demand ~~Four Shillings~~	—	—	—	—	4	0		MELDRETH PM 19 DE 25
		Interest 1925	—	—	4	2	—	—	Balance	
		Total	2	4	7	2	4	7		
1926		Balance	2	—	4					
June	3	on Demand				1	0	0		MELDRETH AM 3 JU 26
June	26	on Demand	—	—	—	1	0	0		MELDRETH AM 26 JU 26
		Carried forward ... £	2	—	7	2				

CANCELLED

CANCELLED

Security and Identity/Access Management

Mundane Efficiency

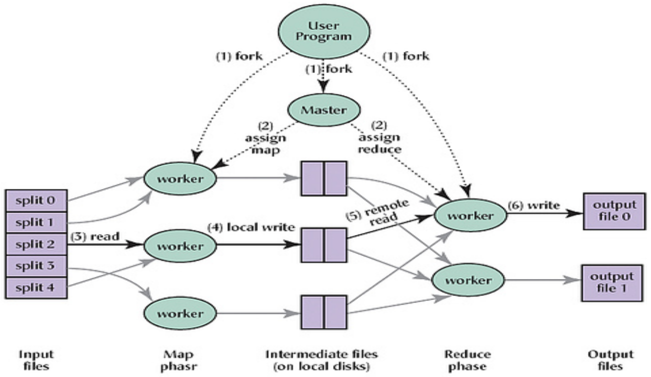
- Support for smart usage: usage reports, billing alerts, cost explorers, life-cycle management, etc.

- Multi-tenant solutions become necessary but take a new way of thinking
  - The cloud is all about elastic resource management. But that’s only part of the solution
  - Also need help with cross-account access management, fine-grained cost/benefit allocation, etc.
  - But things like IAM, detailed billing reports, and Cost Explorer can help.

- Ditto for reusability of services, computations, data, results, etc.

From Pioneers to Settlers

From "do it exactly the way I tell you to" to "just do it"



```
foreach inputDataSet in sources
    computeAlgorithm
```

Production readiness has to be built in

The potential of Big Data

The production use of Big Data

- Automate “everything”
- Security by default
- End-to-end change journals and audit logs
- Serverless, event-driven computing
- Reproducibility, support for data amendments
- Support for relentless cost reduction

Luckily the Cloud can do most of the heavy lifting for you

- CloudFormation, Opsworks, Spark on EMR
- VPC, IAM, Cloudtrail, AWS Config, KMS, ACM, CloudHSM, Inspector
- CloudWatch Logs & Events, Lambda
- Storage life-cycle management
  - S3, Glacier, Snowball, DirectConnect
- Tags, usage reports, billing alerts, cost allocation, detailed billing reports, cost explorers, Trusted Advisor, QuickSight, Redshift

THANK YOU.

Spark summit

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO

Images used in this talk

- Slide 2:
  - https://pixabay.com/static/uploads/photo/2014/11/20/08/57/success-538725_960_720.jpg
  - https://i.ytimg.com/vi/Nc1-gACc4lo/hqdefault.jpg
- Slide 5:
  - https://upload.wikimedia.org/wikipedia/commons/8/87/IBM_card_storage.NARA.jpg
  - https://upload.wikimedia.org/wikipedia/commons/1/1b/Francois_sagat_in_BLAB_LA_ZOMBIE_by_ARNO_ROCA_%284399970229%29.jpg
  - https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/United_States_one_dollar_bill,_obverse.jpg/1024px-United_States_one_dollar_bill,_obverse.jpg
- Slide 6:
  - https://upload.wikimedia.org/wikipedia/commons/b/b8/Manusingmicroscope.jpg http://images.forbes.com/media/2009/05/04/0504_best-paying-jobs_21.jpg
  - https://cdn.milwaukeetool.com/~/media/Images/Power%20Tools/Corded/6538-21/39789_6538-21v3-lg.jpg
  - https://www.milwaukeetool.ca/fr//cdn.milwaukeetool.com/~/media/Images/Power%20Tools/Corded/6520-21/39679_6520-21v2-lg.jpg
  - https://s-media-cache-ak0.pinimg.com/236x/b1/1a/af/b11aaf393cf23ef3d2baf979d0c69d22.jpg
- Slide 7:
  - http://media.culturemap.com/crop/01/49/633x475/Katy-Freeway-highway-Interstate-10-traffic-traffic-jam-March-2014_165658.jpg
  - https://pixabay.com/static/uploads/photo/2013/07/13/13/41/robot-161367_960_720.png

Images used in this talk (cont.)

- Slide 8:
  - https://upload.wikimedia.org/wikipedia/commons/2/2c/House_destroyed_by_fire_in_village_Burkovo,_Brilyakosky_Selsovet.jpg
  - https://upload.wikimedia.org/wikipedia/commons/b/bf/22_May_2011_Joplin_tornado_damage.jp
  - http://www.meldrethhistory.org.uk/images/uploaded/originals/post_office_book.jpg
  - https://c2.staticflickr.com/8/7410/8981216419_2e53463a33_b.jpg
- Slide 9:
  - https://pixabay.com/static/uploads/photo/2013/07/13/10/24/burglar-157142_960_720.png
  - http://www.cynic.org.uk/holidays/nyc2002/manhattan-s.jpg
- Slide 11:
  - http://www.legendsofamerica.com/photos-americanhistory/William-becknell.jpg
  - http://1.bp.blogspot.com/_bvtos4pi3no/S-BtPUuxDkI/AAAAAAAAA0E/RnSb1I7NWPY/s1600/DSCN1071.JPG
  - https://pixabay.com/static/uploads/photo/2015/04/24/22/47/san-francisco-738416_960_720.jpg
  - https://c3.staticflickr.com/3/2133/2179187226_e2e107e0cd.jpg
- Slide 12:
  - http://www.mountainphotography.com/images/275/20051222-Timpanogos-Sage.jpg
  - http://www.systemsolutionsdevelopment.com/wp-content/uploads/2015/05/chicagoatnight.jpg