

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?

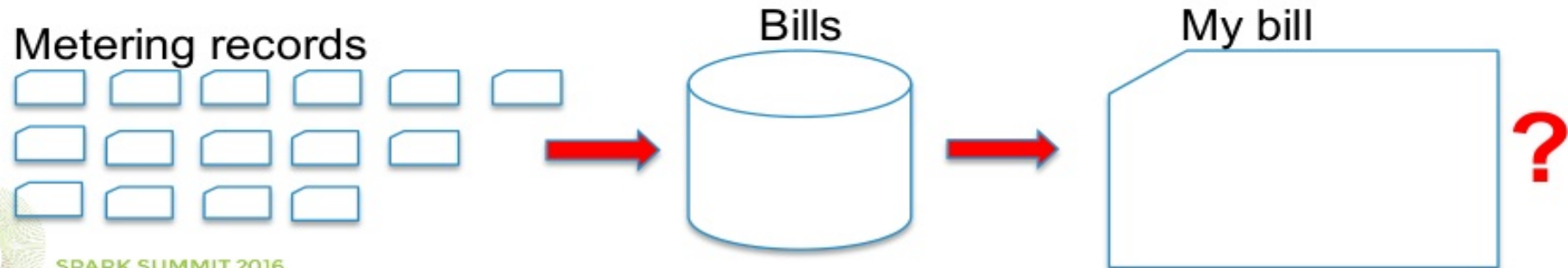
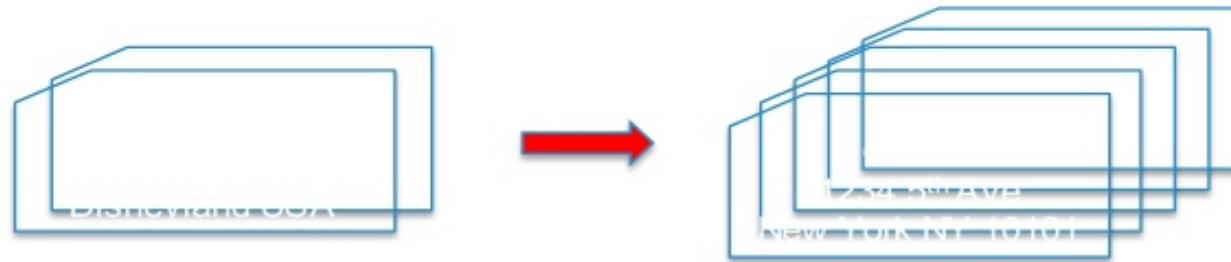


Production “ilities”

- **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



Algorithmic efficiency → “mundane” efficiency

Constants matter



The never-ending battle against waste



Fine-Grained Control vs. Ease-of-Use



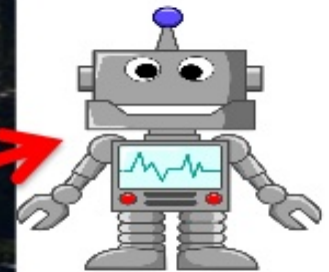
VS.



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



Scale → Automation

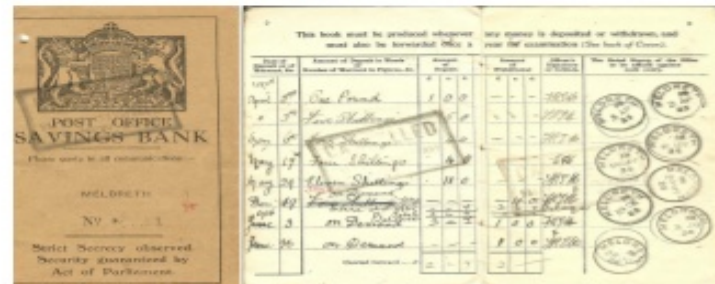
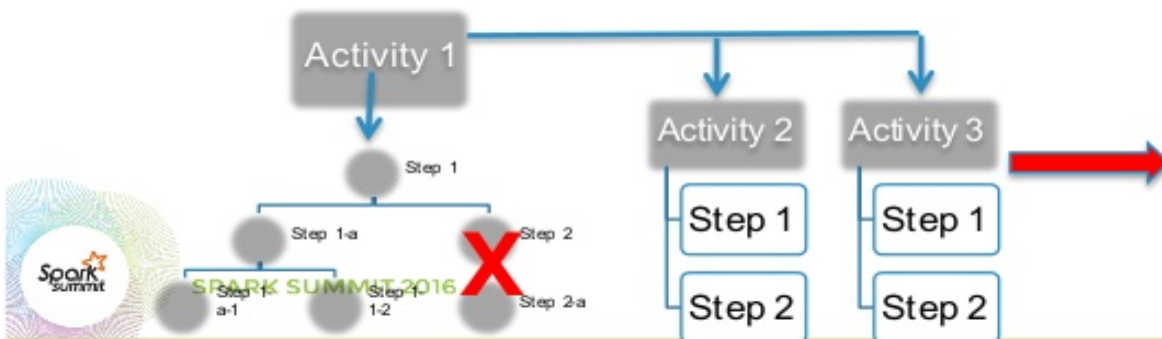


High Availability

Single points of failure vs. “large scale events”



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



Security and Identity/Access Management



SPARK SUMMIT 2016

Mundane Efficiency

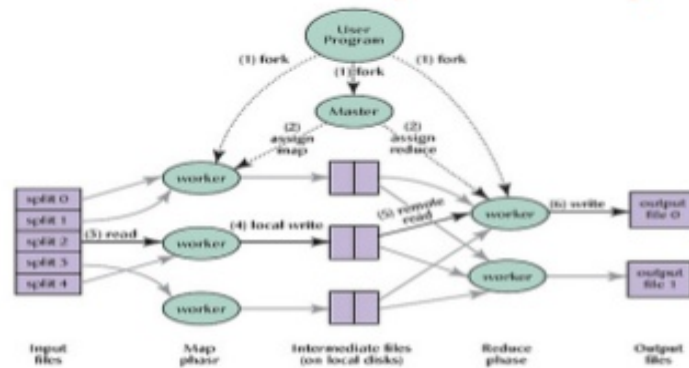
- **Support for smart usage:** usage reports, billing alerts, cost explorers, life-cycle management, etc.
- **Multi-tenant solutions become necessary but take a new way of thinking**
 - The cloud is all about elastic resource management. But that's only part of the solution
 - Also need help with cross-account 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”***



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 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-gACc4Io/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/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](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.milwaukeejournal.com/~media/Images/Power%20Tools/Corded/6538-21/39789_6538-21v3-lg.jpg
 - https://www.milwaukeejournal.com/~media/Images/Power%20Tools/Corded/6520-21/39679_6520-21v2-lg.jpg
 - <https://s-media-cache-ak0.pinimg.com/236x/b1/1a/af/b11aaf393df23ef3d2baf979d0c69d22.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 Sel sovet.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>

