RESOURCE SHARING

Mark Ryland
Chief Solutions Architect
Worldwide Public Sector Team

amazon web services

SHARING: TWO DIMENSIONS

- Technology and Capability
- Economics / Costs

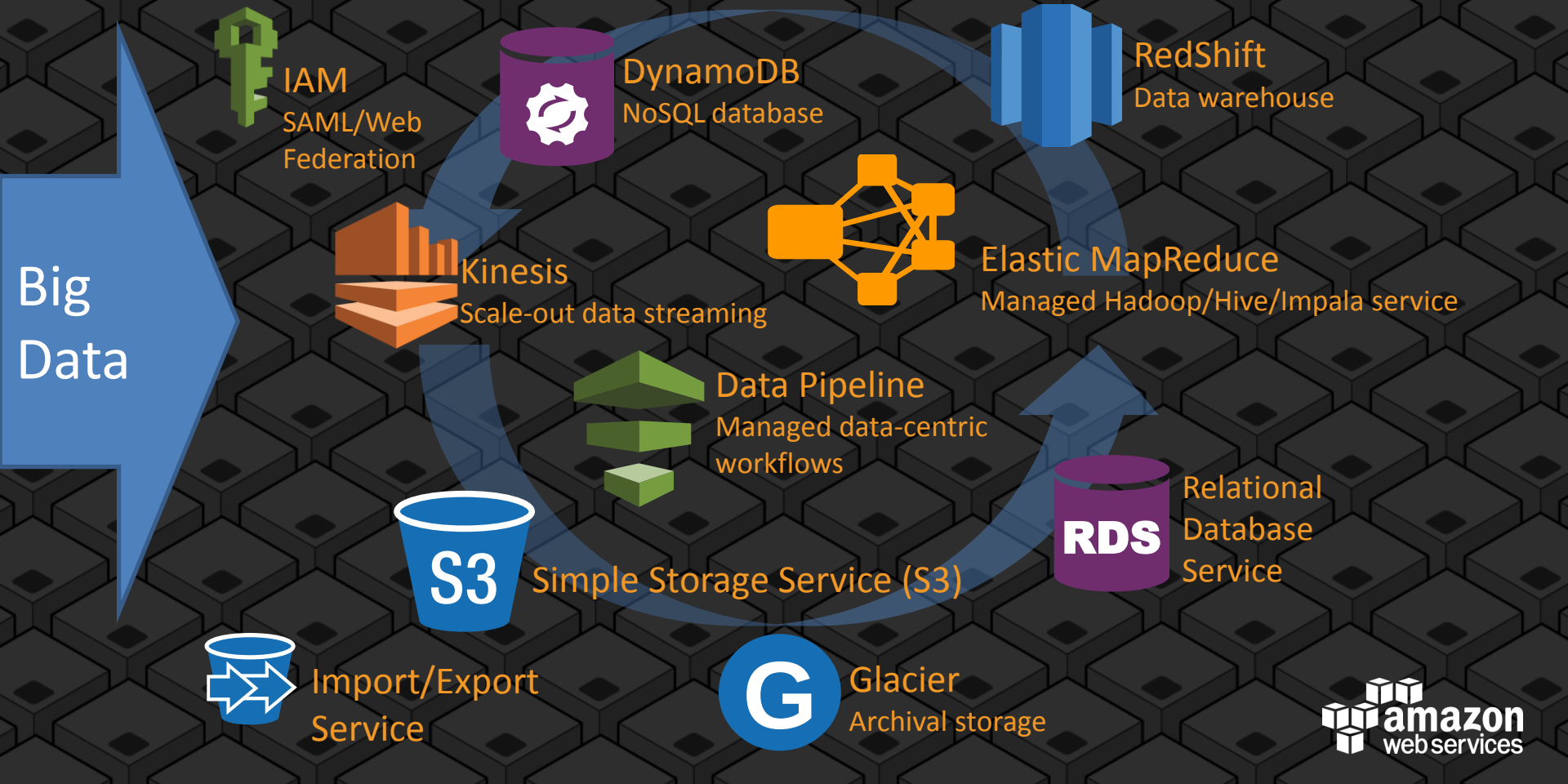
DATA

- Public datasets – free data of broad interest
  - AWS: 1000 Genomes, NEX, Common Crawl*
  - Yahoo! Webscope (larger datasets)
- S3 requester-pays for cost sharing
- Community/private: sharing + rich access control
  - Including Web Identity Federation*
- Data transfer services* (contra data gravity)

COMPUTE

- Pre-baked AMIs with tools (and data)
  - E.g., BioLinux; NITRC tools
- AWS Marketplace (aws.amazon.com/marketplace)*
- Not limited to single compute nodes…
  - Open source cluster management / devops tools
  - Cloud Formation
- Spot pricing*

MANAGED BIG DATA SERVICES



ECONOMICS

- Grants program (free)
- Spot pricing (cheap computation)
- Volume discounting
- Institutional / co-op pricing

GRANTS PROGRAM

- Student, teaching, and research; e.g.,
  - AMP Lap Spark/Shark (and related) stack
  - University of Oxford malaria project
  - UCSD assistive computer vision technology
  - UCSF analytics program
  - Hundreds of others…

EDUCATION AND TRAINING

- Free videos and papers*
- Free hands-on labs*
- Paid big data training course (new)
- Working with community to aggregate open source materials and curricula

THANK YOU!

APPENDIX

*REFERENCES

- Public datasets: http://aws.amazon.com/publicdatasets/
- SAML, Web Identity Federation: http://aws.amazon.com/iam
- Cloud data transfer service: https://www.globus.org/
- AWS Marketplace: http://aws.amazon.com/marketplace
- Spot pricing: http://aws.amazon.com/ec2/purchasing-options/spot-instances/
- Instructional videos and hands-on labs:
  http://aws.amazon.com/training/intro_series/
  http://aws.amazon.com/training/self-paced-labs/
- New big data course: http://aws.amazon.com/training/course-descriptions/bigdata/

Cloud and Big Data: Natural Affinity

- Big data:
  - Variety, volume and velocity requiring new tools
  - Potentially massive datasets, future size unknown
  - Iterative, experimental style of data manipulation and analysis
  - Frequently not a steady-state workload; peaks and valleys
  - Absolute performance not as critical as “time to results”; shared resources (*e.g.*, single large cluster) = bottleneck
- Utility (true cloud) computing:
  - Variety of compute, storage, and networking options
  - Massive, virtually unlimited pay-as-you-go capacity
  - Iterative, experimental style of infrastructure deployment/usage
  - Most cost-efficient with variable workloads
  - Parallel compute projects give autonomy to workgroups, they get faster results
  - **Cloud democratizes big data**