From Raw Data to Usable Results, Build, Rinse and Repeat

Asya Shklyar
SCALE 2017

Data Sources

- Sequencer
- Sensor, i.e. Temperature, Light, Speed
- Camera
- Microscope
- Microphone
- Camera, image or video
- Human input
- GPS
- Telescope
- Radar

Data Formats

- Text
- Structured text
- Table or Spreadsheet
- Database or structured data
- NOSQL
- Columns
- Queue
- Graphics
- Video
- Compressed

Data Caching

- Local instrument disk
- Local fast cache, connected by a fast network
- Distributed cache
- RAM

Data Transfer

- Network connection
- Internal vs External
- Public Internet vs Internet2 vs a private WAN
- Transfer tools (schedule, resume, throttle, checksum, compress, uncompress, tag, notify, policy, security, multi-threaded/parallelized)
- Data size

Data Integrity

- Checksumming
- Small files vs Large files
- Encryption
- Public keys
- Key management
- Confirmation

Data Security

- Encryption tecniques
- On the fly
- Real time
- Firewalls
- Science DMZ concept
- ACLs

Data Storage

- Flash vs DRAM vs Rotating Disk
- Connectivity, 1 GB, 10 GB, 40 GB, 100 GB, 200 GB, 400 GB
- Ethernet vs Infiniband
- How many ports on either side
- Throughput
- Blocking vs Non-blocking
- Distributed
- Object Storage
- S3 vs Swift
- Storage Layers

File Systems

- Linux ext3, ext4, xfs
- Parallel, GPFS, Lustre
- NFS
- GlusterFS
- pNFS
- Fuse
-

Data Management

- Data Policies and Rules
- iRODS
- Robinhood
- StarFish
- Built-in (DDN, Avere, CleverSafe)

Metadata Management

- generation
- transfer
- modification
- attribute change

Workflows

- Data normalization
- Data extraction
- Data import
- Data export
- Data reduction
- Enrichments
- Combination
- Sort
- Search
- ML
- Bioinformatics specific (SnakeMake etc)

Data-centric infrastructure

- Hardware
- GPUs, FPGAs, ASICs
- Software
-

Data Collaboration

- Cloud(s)
- Email, DropBox, Google Drive, Basecamp, Box
- scp, rsync, parsync, Globus/Aspera,

Data Archival

- Cloud vs Cheap Disk vs Tape
- Retention
- Retrieval (takes time)
- Cost

Data People

- Data Scientist
- Chief Data Officer
- Compliance Officer
- Backup and Recovery Personnel
- DBAs
- People that collect and sift through logs
- Information Security Officers