Hadoop

• Platform for Big Data
  – Very large Datasets
  – Horizontally Scaling computation

• Platform for Applications
  – Apache Spark
  – Hbase
  – Hive
  – More
Your Very Own Cluster

• Personal cluster
  – Allocated for the duration of the job
  – Size according to needs
    • Uses RSM nodes
    • 8TB disk space per node
    • 2x14-core CPUs per node
    • 128 GB RAM per node
  – Group access to cluster
    • Login via ssh
Monitoring

- Hadoop web interfaces
  - Available via ssh tunnel to nodes
  - DFS manager
  - Job Manager
Software

• Apache Hadoop 2.7.2
• Apache Spark 1.6.0
• Apache Hbase 1.1.3
• Other packages available by request
Persistence

- Nodes reserved for private use
  - (Meter is running while idle)
- Contents of HDFS remain until job is finished
- User must extract all data they wish to save
- Nodes will be wiped after reservation
Example Cluster

• 12 Node Test Cluster Aggregate Stats:
  – 336 cores
  – 1536 GB RAM (~1.5TB)
  – ~120 TB Disk (~60 TB HDFS with dual replication)

• All the Nodes (768 node cluster) Aggregate Stats:
  – 21504 cores
  – ~96 TB RAM
  – >6 Petabytes storage
Questions?