NIST Big Data Public Working Group

Big Data Future Trends

- Research Data Alliance Big Data Analytics IG

Chairs
Kwo-Sen Kuo, Bayesics
Morris Riedel, fz-juelich.de
Peter Baumann, jacobs-university.de

Kwo-Sen Kuo, Ph.D.

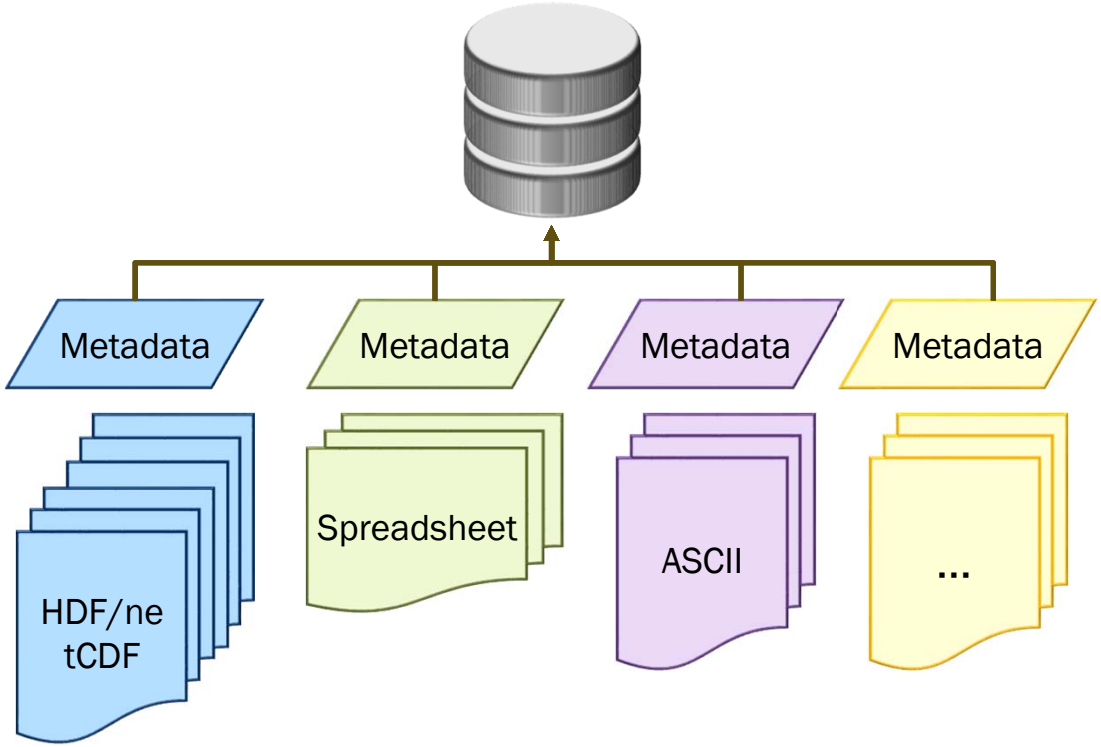
- C*x*O and founder of Bayesics, LLC, where *x* is a variable.
- Also affiliated with
  - NASA Goddard Space Flight Center, Greenbelt, Maryland, and
  - University of Maryland, College Park, as a visiting scientist.
- *Jack* of the *atmospheric remote sensing* and *data information system/technology* trades, among others.
- Interested in crafting *Power Tools* for scientists to focus on science and conduct research more effectively.
- Co-investigator of a number of NASA-funded projects.

Overview

Scientific data are our focus!

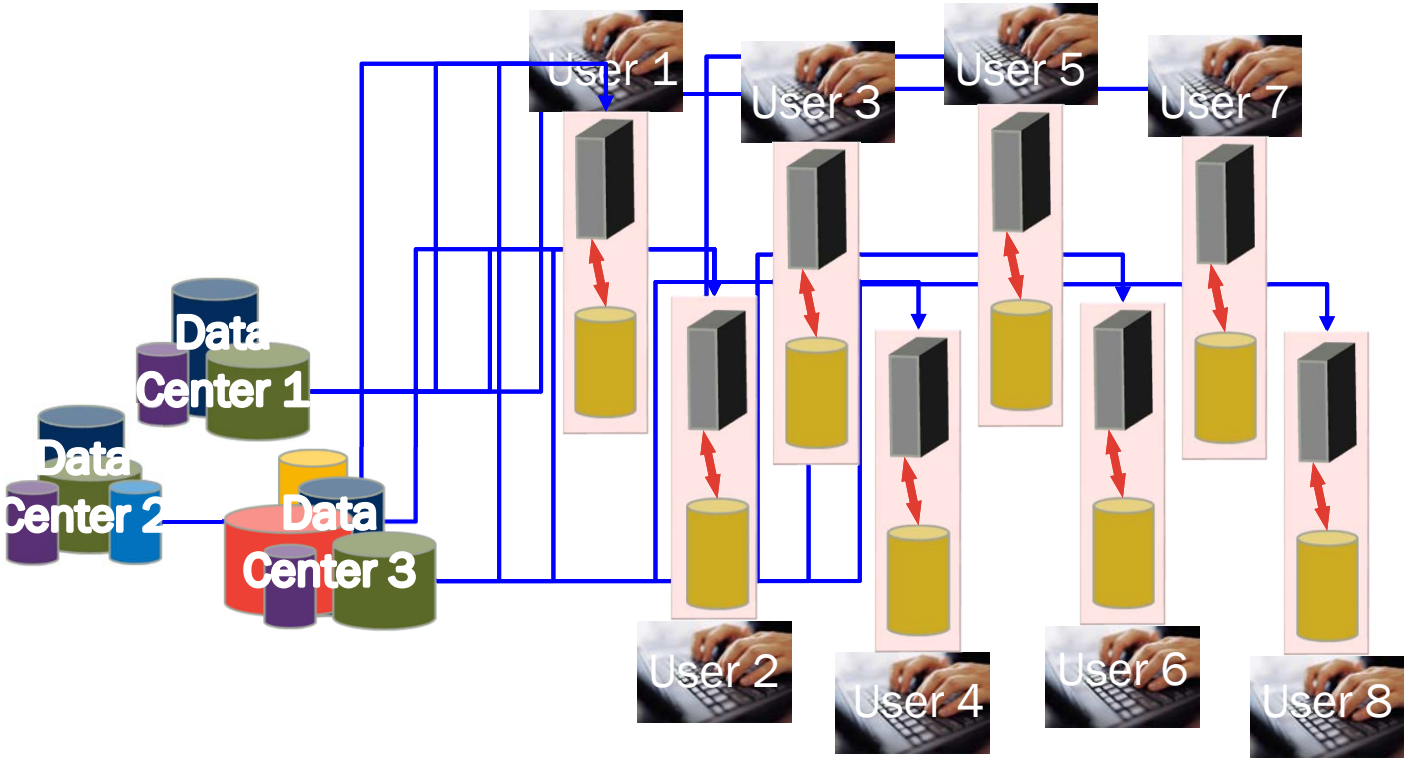
- Current Data Archive and Distribution Practices
  - Root of all ills…
- Current Data Analysis Practice
  - Inevitable results of current data archive and distribution practices
- Traditional High Performance Computing (HPC) Architecture
  - Suitable for tightly coupled, *e.g.* MPI, parallelism
- Data Intensive Architecture (*a.k.a* shared-nothing architecture)
  - Suitable especially for "embarrassingly parallel" problems
- Data Archive, Distribution, Analytics – All-in-one
  - Should allow us to do previously difficult or nearly impossible things!

Current Data Archive and Distribution Practices



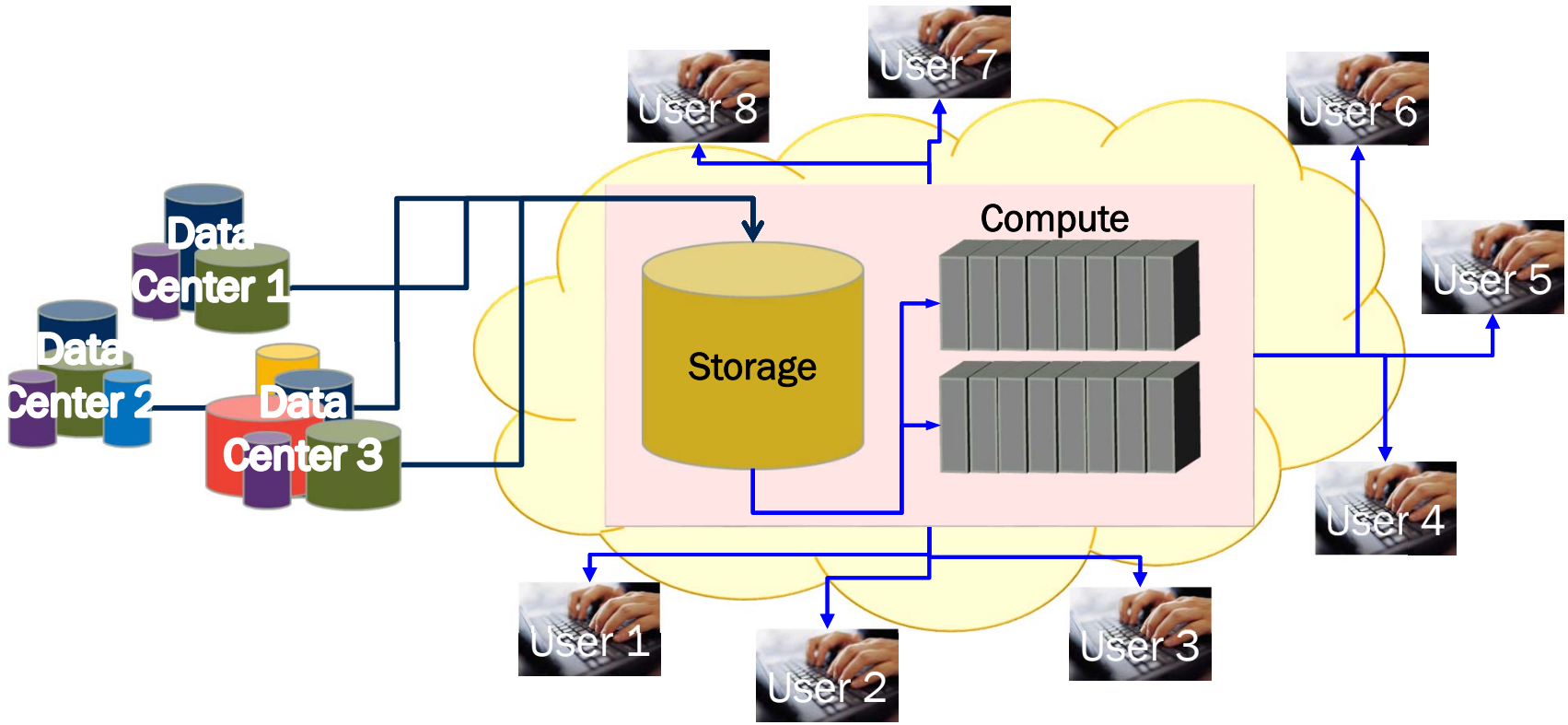
- File based archive distribution.
- Only metadata are cataloged in databases, thus searchable.

Current Data Analysis Practice



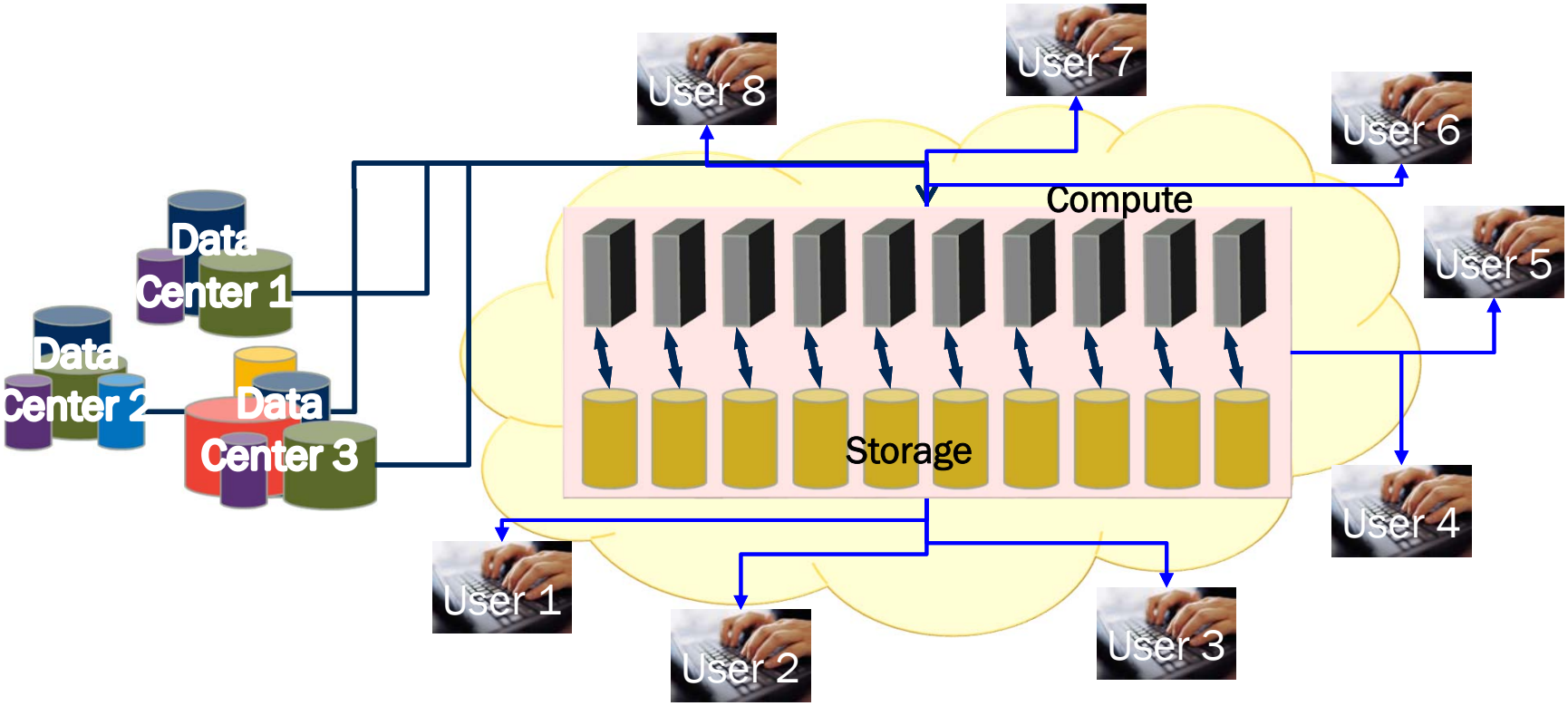
- Everyone downloads needed data from data centers.
- Data analysis is conducted on local resources, mostly serially.

Traditional HPC Architecture



- Fast, central, large storage system.
- Suitable for MPI parallelism, but few scientists can program with it.

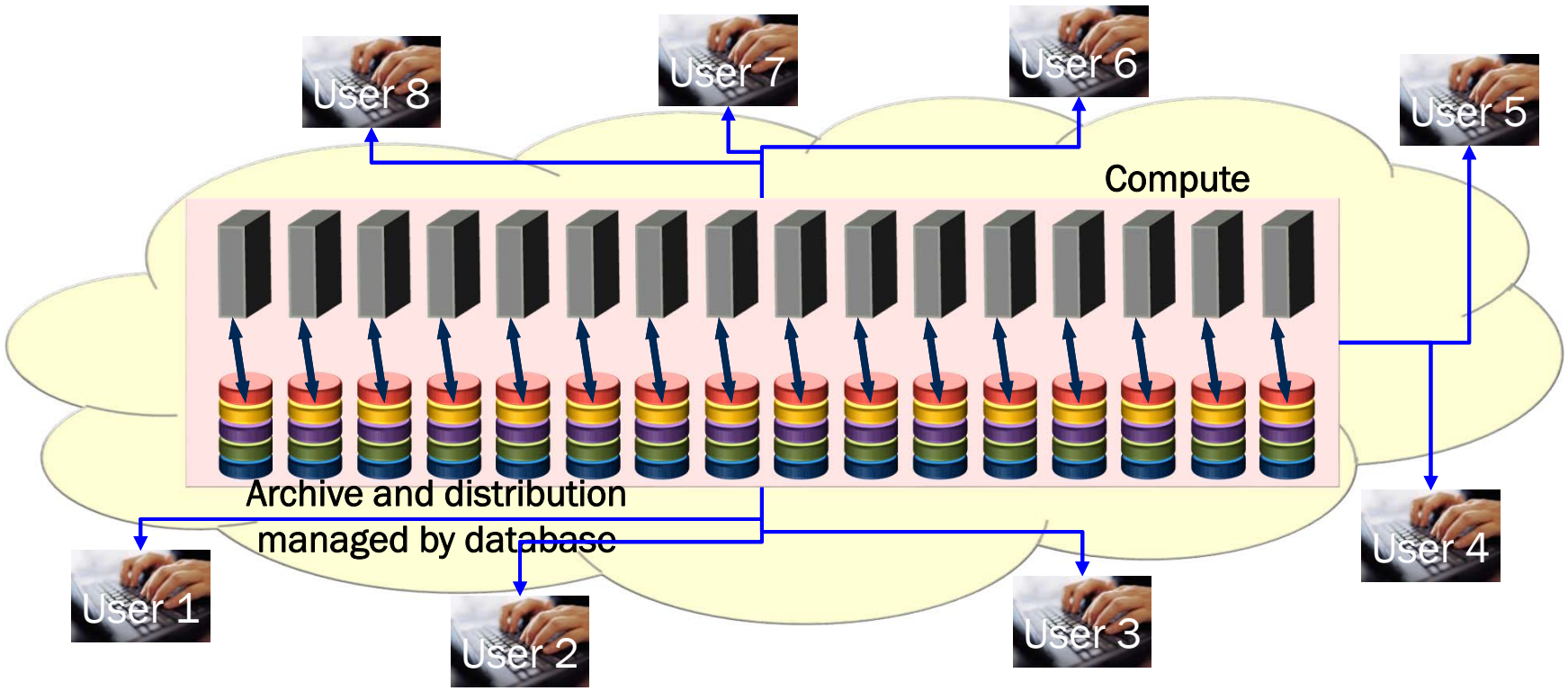
Data Intensive Architecture



- Distributed storage with local compute access.
- Loosely coupled parallelism requires little or no inter-process communication

RDA Research Data Sharing without barriers

Data Archive, Distribution, and Analytics All Together



- Data directly "archived" into parallel databases for distribution and or analysis.