High Performance Computing with Confidential Data

Tim Mulcahy, Principal Research Scientist
Daniel Gwynne, Chief Engineer
Johannes Huessy, Principal Research Analyst

NORC at the UNIVERSITY of CHICAGO

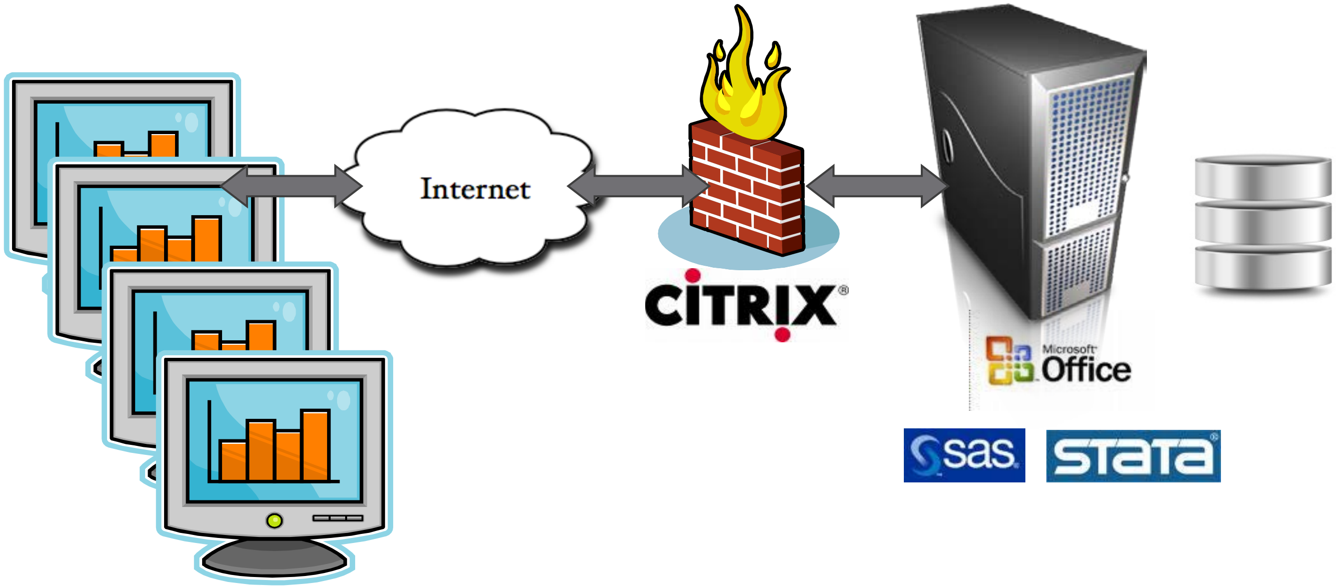
Overview

- History of the NORC Data Enclave
- Emerging Challenges with Research Data
- Technical Issues
- Solutions
- Forthcoming Developments

What is the Enclave?

NORC at the UNIVERSITY of CHICAGO

The Enclave is an environment that allows for secure remote access to confidential microdata.



Through the use of a secure terminal session, researchers analyze sensitive data without the data ever leaving the FISMA compliant secure data center.

History of the NORC Data Enclave

NORC at the UNIVERSITY of CHICAGO

- Founding Sponsor: National Institute on Standards and Technology (NIST)
  - Surveys from the Technology Innovation Program (TIP) grantees
- Other early adopters:
  - USDA Economic Research Service: Agricultural Resource Management Survey
  - Kauffman Foundation: Kauffman Firm Survey
- At its inception, the Enclave was focused on survey data, most of it economics-related.

Select Current Sponsors

- Financial Transaction Data
  - Private Capital Research Institute
  - Financial Crisis Inquiry Commission
- Large Scale Health Data
  - CMS
  - Maine Health Data Organization
- Other Federal Statistical Data
  - NSF
  - USDA NASS
- Private Sector Survey Data
  - Annie E Casey Foundation

Emerging Challenges

Why Big Data and High Performance Computing at NORC

NORC’s strategic imperative is to grow our expertise and capabilities in the management, analysis, and dissemination of administrative and other data; this will require the ability to run programs to repurpose and statistically analyze terabytes of data.

From the Bureau of Labor

“What is the future of the use of Big Data for the U.S. statistical system? I see one immediate potential: the use of Big Data to improve the quality of our estimates within our current methodological frameworks. This may include studies of comparability between official and Big Data–derived estimates, the use of Big Data for modeling and imputation, and—in some cases—the use of Big Data for direct estimation.”

Bureau of Labor Statistics Associate Commissioner Michael W. Horrigan

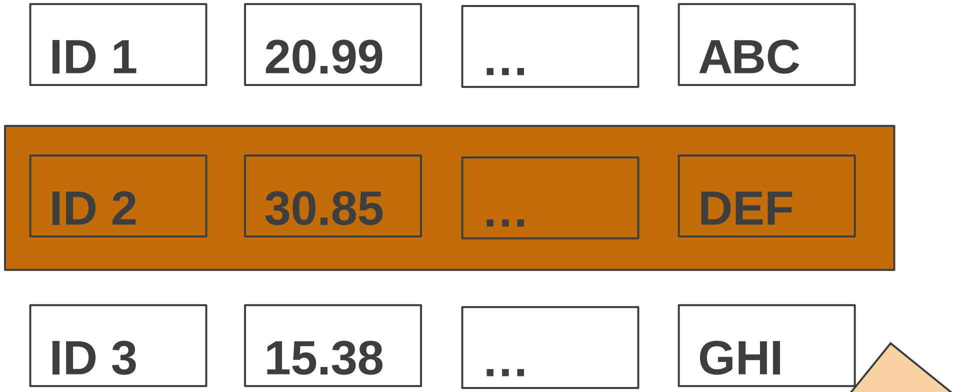
Emerging Challenges

- Fields that have traditionally worked with survey data are increasingly turning to transaction, record and machine-generated data to create policy-relevant research products
- These new data types are typically several orders of magnitude larger and are often not structured in a manner that make them easily analyzed by researchers

Technical Issues

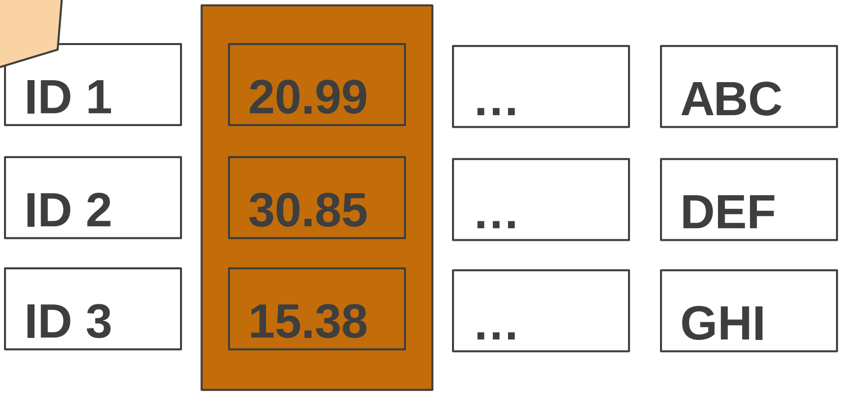
- Different organizations have different needs and priorities for big data
- In contrast to commercial organizations or social media networks that need to load and update data continuously, research organizations with relatively static datasets and frequent complex queries need a strategy for big data that focuses on reducing query execution time

Column Orientation



ID 1	20.99	…	ABC
ID 2	30.85	…	DEF
ID 3	15.38	…	GHI

Column storage reduces the amount of data that needs to be read by focusing only on the relevant columns

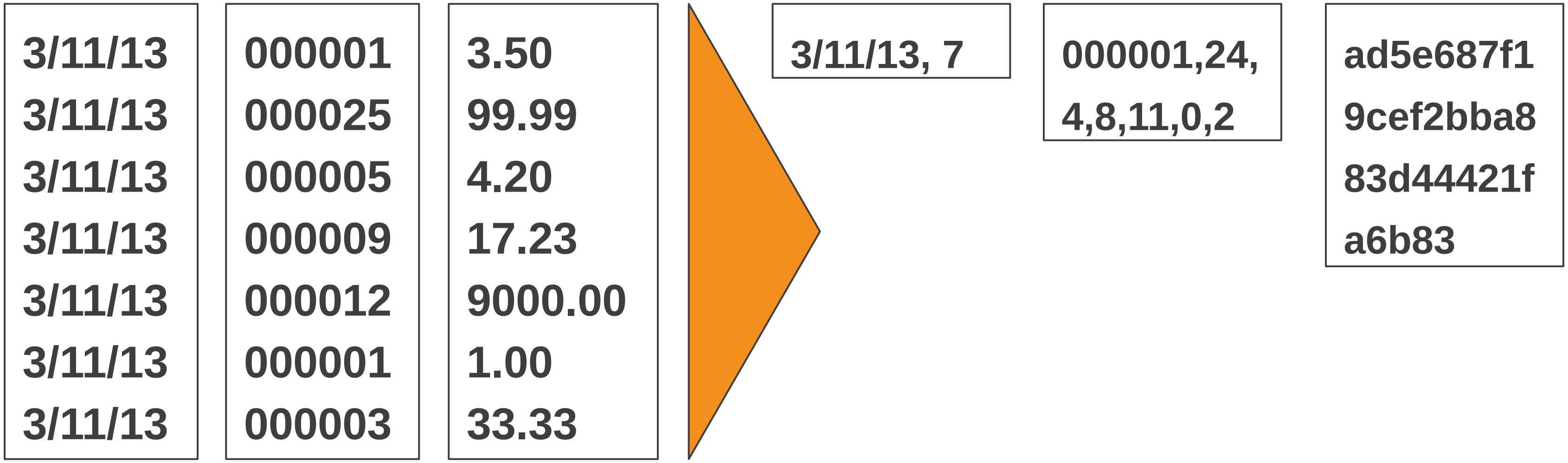


ID 1	20.99	…	ABC
ID 2	30.85	…	DEF
ID 3	15.38	…	GHI

For example, if one wanted to return a column average from a 100 million record file with 1000 columns each storing ten bytes per record, row oriented databases would read 1 TB of data while column oriented databases would only read 1GB

Compression

- The simplest response to high volumes of data is to make it smaller
- An ideal solution is able to intelligently apply a number of different compression algorithms to each column and then perform queries on the compressed data or rapidly decompress the data for analysis



3/11/13
3/11/13
3/11/13
3/11/13
3/11/13
3/11/13
3/11/13

000001
000025
000005
000009
000012
000001
000003

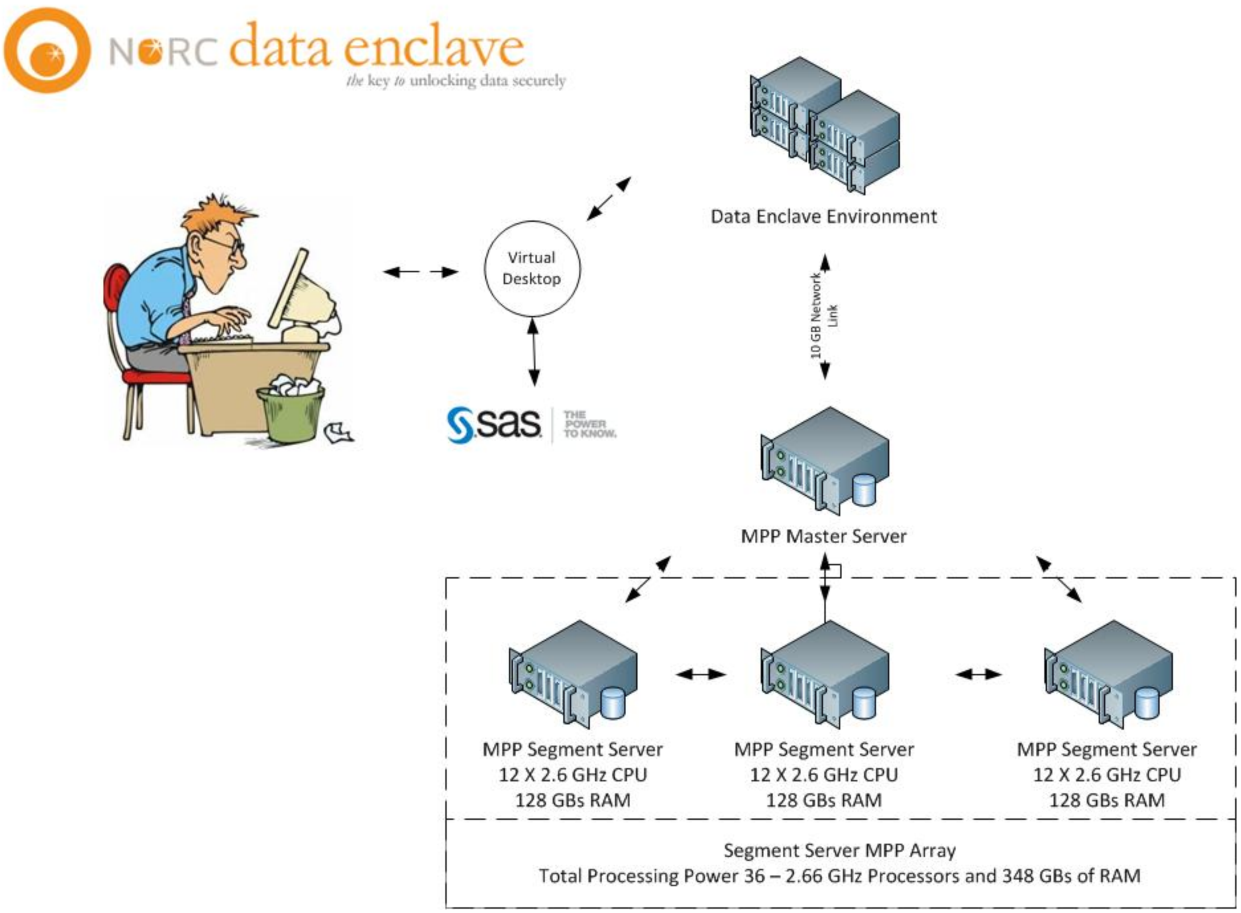
3.50
99.99
4.20
17.23
9000.00
1.00
33.33

3/11/13, 7

000001,24,
4,8,11,0,2

ad5e687f1
9cef2bba8
83d44421f
a6b83

Parallel Processing



Unstructured Data

- Much of the emerging data for social science research is not conveniently organized into tables or coded for easy discovery
- Harnessing these datasets means integrating data tables with unstructured files and programs for mining structured information out of those files
- In practice this means integrating with Hadoop leveraging HDFS technology for non-tabular data

Performance

- A solution that implements these technologies can reduce query execution times by an order of magnitude in comparison with traditional databases and by a factor of a thousand in comparison with queries against flat files
- This makes research projects possible in timeframes that would otherwise have been unrealistic or unthinkable.

Forthcoming Developments

- We support big data, now what?
  - How do we support better application integration to make large datasets transparent to more people?
  - How do we support the integration of disparate real-time data streams into harmonized data files?
  - How do we leverage large confidential datasets to produce both anonymized and useful public data products?