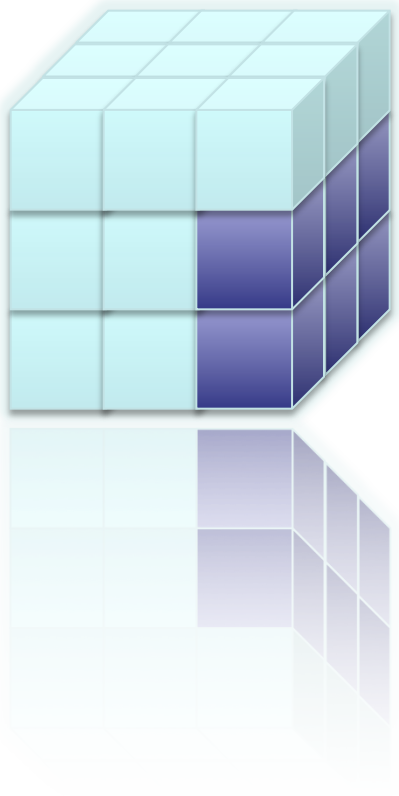


Actionable Requirements for Big Science Data Management

Paul G. Brown

U. Wash Summer Institute - 2010

Zetics

Big Data. Big Analytics. Big Egos.

Topics

- Science Use Cases with a Big Data Flavor
- SciDB – Architectural Features
- Project Status

Science Groups

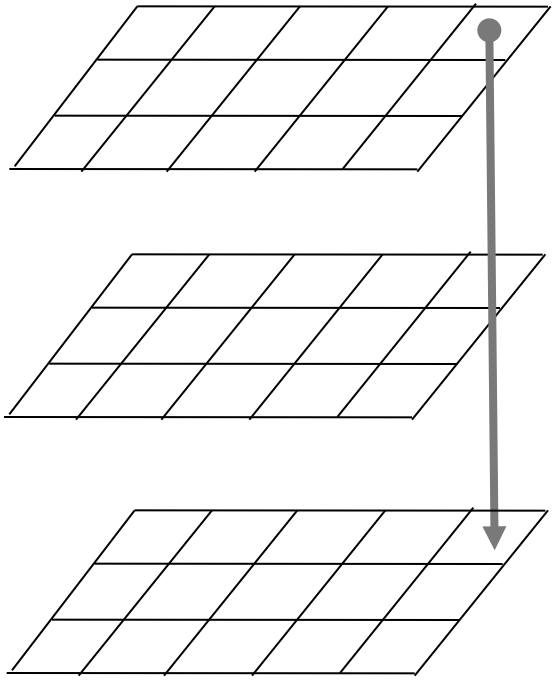
- Astronomy – 'pointing the camera up' (LSST – via. SLAC)
- Remote Sensing – 'pointing the camera down'

Technical Lessons

- Arrays as Storage Model
- Extensibility
    - Data Types and Scalar Functions
    - Operators in the Array Algebra

Common Operations

- Images are really 'Arrays' – Pixel values are really 'data'
- Super-imposition of 'Images'



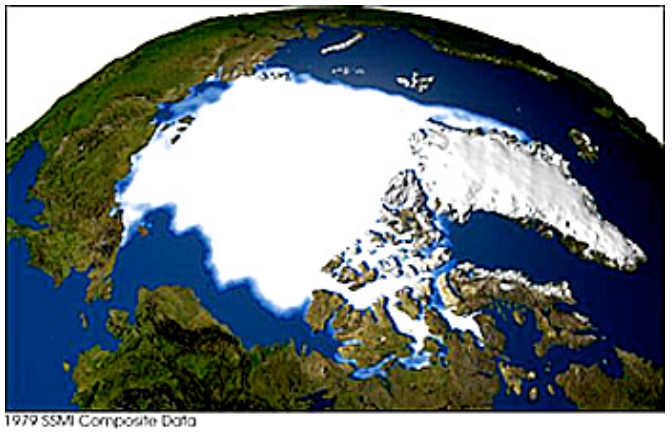
A [i, j]

+

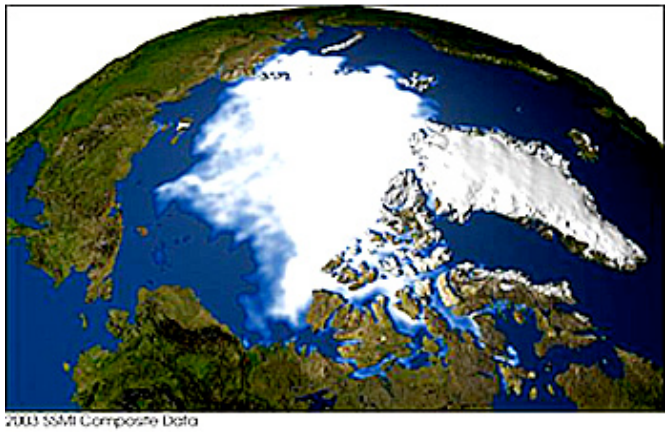
B [i, j]

=

C [i, j]



1979 SSMI Composite Data



2003 SSMI Composite Data

Un-common Operations

- Science Data – Lots of Numbers
  - Lots of DOUBLE and INT, few DECIMAL
  - Exotics – User-Defined Types: COMPLEX, RATIONAL
  - Random Variables – N (x, v)
- Missing Information – More than NULL
  - Array cells can be 'EMPTY'
  - Array cell values can be 'MISSING', or 'Out of Bounds'
- New Operations
  - Consider : extensible relational operators (beyond Proj, Rest, Join, Union, etc).
  - FFT (input : Array) -> output : Array
  - Feature_Detect (input : Array) -> output : Array < Array >
  - Iterative Methods: do (init(), iter(), until ())

SciDB Example

1. Use arrays as the logical building block of the data model.

```
CREATE ARRAY Observations
  < V: Double > [ I=0:3,3,1, J=0:3,3,1];
```

2. Use as a unit of physical storage, and as processing 'chunk'.

	0	1	2	3
0	0.02	0.01	0.01	0.02
1	0.01	0.01	0.5	0.02
2	0.01	0.02	0.01	0.01
3	0.02	0.01	0.02	0.02

Chunk 1

SciDB Example

1. Use arrays as the logical building block of the data model.

```
CREATE ARRAY Observations
  < V: Double > [ I=0:3,3,1, J=0:3,3,1];
```

2. Use a unit of physical storage, and as processing 'chunk'.

	0	1	2	3
0	0.02	0.01	0.01	0.02
1	0.01	0.01	0.5	0.02
2	0.01	0.02	0.01	0.01
3	0.02	0.01	0.02	0.02

Chunk 2

SciDB Example

1. Use arrays as the logical building block of the data model.

```
CREATE ARRAY Observations
  < V: Double > [ I=0:3,3,1, J=0:3,3,1];
```

2. Use a unit of physical storage, and as processing 'chunk'.

	0	1	2	3
0	0.02	0.01	0.01	0.02
1	0.01	0.01	0.5	0.02
2	0.01	0.02	0.01	0.01
3	0.02	0.01	0.02	0.02

Chunk 3

SciDB Example

1. Use arrays as the logical building block of the data model.

```
CREATE ARRAY Observations
  < V: Double > [ I=0:3,3,1, J=0:3,3,1];
```

2. Use a unit of physical storage, and as processing 'chunk'.

	0	1	2	3
0	0.02	0.01	0.01	0.02
1	0.01	0.01	0.5	0.02
2	0.01	0.02	0.01	0.01
3	0.02	0.01	0.02	0.02

Chunk 4

The 'Provocative Assertions' Slide

- Traditional DBMSs have the wrong data model
  - Tables are impossibly slow at simulating arrays
- SQL has the wrong operations
  - Need to regrid and cluster, not join
  - Need analytical operations like covariance, clustering, et al
- SQL is missing needed features
  - Uncertainty
  - Provenance (lineage)
  - Versions
  - No overwrite

Project Development Status

- **Development underway for 2 years**
  - **Project initially driven by science community**
  - **Team of 20+ volunteers from academic and science communities**
  - **Threw away V (0.0 – 1.0 x i).**
- **Proof-of-concept demos and projects**
  - **Public demo at VLDB and XLDB-3 in August '09**
  - **3 POC's in quantitative finance, genomic sequencing, sky survey data**
- **Will be open source core with an enterprise version offering support and additional functionality**
  - Open source in order to foster a community of contributors and to insure that data is never "locked up"