Agrios: A Hybrid Approach to Big Array Analytics
What kind of tool should we use for analyzing Big Data?

Make analytic systems work better with large datasets

Create database systems that perform sophisticated analytics

Combine an analytic system with a database: “let each do what it does best”
One example of a hybrid system: Agrios

- Integrates R and SciDB

- Middleware between the two systems
- Powerful analysis system and programming language
- Vectors and arrays are fundamental data objects
- High-level operations on data
- Open-source

```
res <- A %%*% B
```

**R**

**Operations**
- +
- \%
- *
- /
- ^
- .
- &&

**Data**
- A
- C

**SciDB**

**Operations**
- apply()
- mult()
- join()

**Data**
- B
- D
- E
Array database management system

Arrays are the fundamental data objects

Low-level operations on arrays

Scales well

Open-source

store(multiply(A,B), res);
Agrios architecture and workflow

R script

Parser

Accumulator

Stager

Executor

R object or plot

R

SciDB
Hybrid systems are not without problems.

At both hybrid components:

- Data is stored
- Analytic work is performed on the data

This means data must move between components.

We need to decide *when* data should be moved between hybrid components.
Decisions about data movement matter:

Vectors $R$ and $C$ are stored at $B$, and we need their product at $A$.

Upshot: There are choices when it comes to data movement, and some choices are better than others.
This decision amounts to filling in the question marks with operation execution locations

\[
\begin{align*}
operation_1 &: R \\
operation_2 &: SciDB \\
operation_3 &: SciDB \\
\cdots \\
operation_n &: R
\end{align*}
\]

This yields a \textit{staging}
Some possible stagings:
	here are many more …
Select the best staging

What is the best staging?
How do we find the best staging?
Agrios prototype system

- Several key operations currently implemented
  - Matrix multiplication
  - Subscript
  - Aggregate sum
  - Binary arithmetic

- Transformations currently implemented
  - Subscript-through-binary addition
  - Commute
  - Associate
  - Subscript-through-matrix multiplication

- Instrumented for experiments
Original express...
For execution

R script

Original expression

Equivalent logical expressions

Equivalent physical expressions

1000

10

30,000

750
Strategy 1 – simple staging

Compare automated staging to:

- “All-at-R”
- “All-at-SciDB”

![Graphs comparing elements transferred]
Rewrite plans, then stage

- Creates more opportunities for optimization
- but uses more system resources
Accumulate expressions prior to staging and expression rewriting
Agrios is a hybrid system integrating R and SciDB
- R scripts as inputs
- Automatically reduces data movement between R and SciDB
- Outputs results in R

Next steps & active questions
- Exploring performance characteristics of Agrios’ optimizer
- Better understanding plan space
- Implementing additional operators and transformations
- Exploring interactions between rules
Thank you.

This research is supported by National Science Foundation Grant #1110917 and Intel’s Science and Technology Center for Big Data.