Example

LOAD

```
A = LOAD ‘traffic.dat’ AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
            COUNT(A);
D = FILTER C BY ip IS ‘192.168.0.1’
            OR ip IS ‘192.168.0.0’;
STORE D INTO ‘local_traffic.dat’;
```

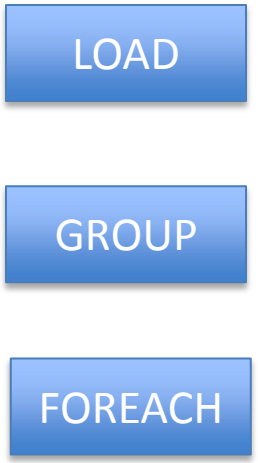

Adapted from slides by Oliver Kennedy, U of Buffalo

Example

```
A = LOAD ‘traffic.dat’ AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
            COUNT(A);
D = FILTER C BY ip IS ‘192.168.0.1’;
            OR ip IS ‘192.168.0.0’;
STORE D INTO ‘local_traffic.dat’;
```

Example

```
A = LOAD 'traffic.dat' AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
            COUNT(A);
D = FILTER C BY ip IS '192.168.0.1';
            OR ip IS '192.168.0.0';
STORE D INTO 'local_traffic.dat';
```



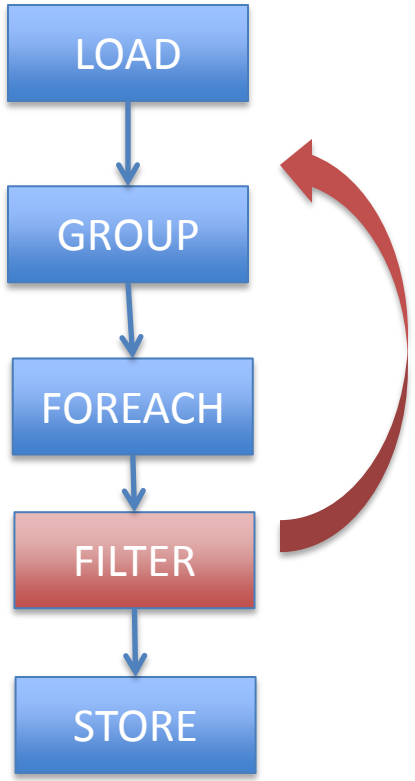
Example

```
A = LOAD ‘traffic.dat’ AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
            COUNT(A);
D = FILTER C BY ip IS ‘192.168.0.1’;
            OR ip IS ‘192.168.0.0’;
STORE D INTO ‘local_traffic.dat’;
```

Adapted from slides by Oliver Kennedy, U of Buffalo

Example

```
A = LOAD 'traffic.dat' AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
                       COUNT(A);
D = FILTER C BY ip IS '192.168.0.1';
                  OR ip IS '192.168.0.0';
STORE D INTO 'local_traffic.dat';
```



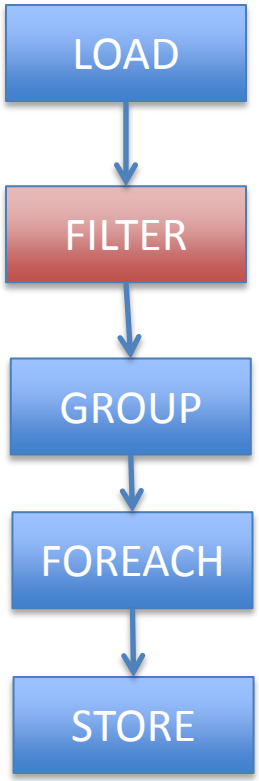
Algebraic Optimization!

Adapted from slides by Oliver Kennedy, U of Buffalo

Example

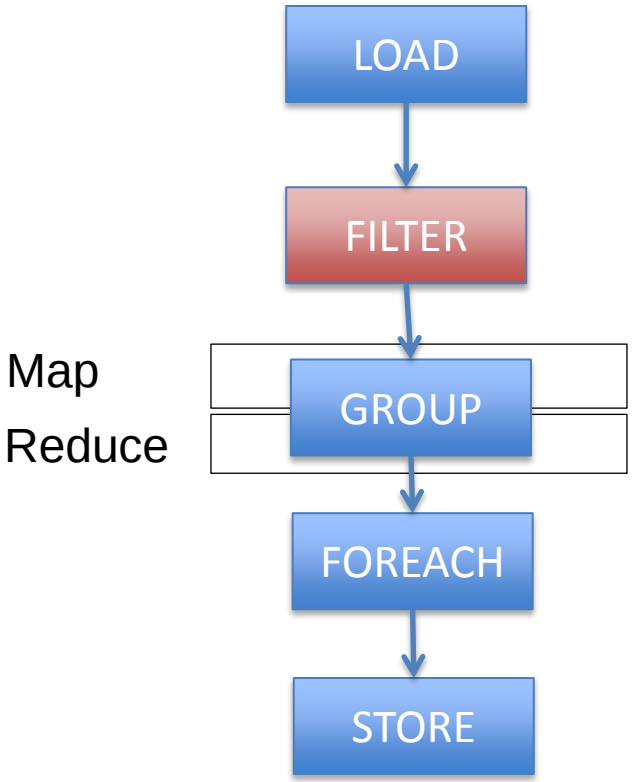
```
A = LOAD 'traffic.dat' AS (ip, time, url);
B = GROUP A BY ip;
C = FOREACH B GENERATE group AS ip,
            COUNT(A);
D = FILTER C BY ip IS '192.168.0.1';
            OR ip IS '192.168.0.0';
STORE D INTO 'local_traffic.dat';
```

Lazy Evaluation:
No work is done
until STORE



Adapted from slides by Oliver Kennedy, U of Buffalo

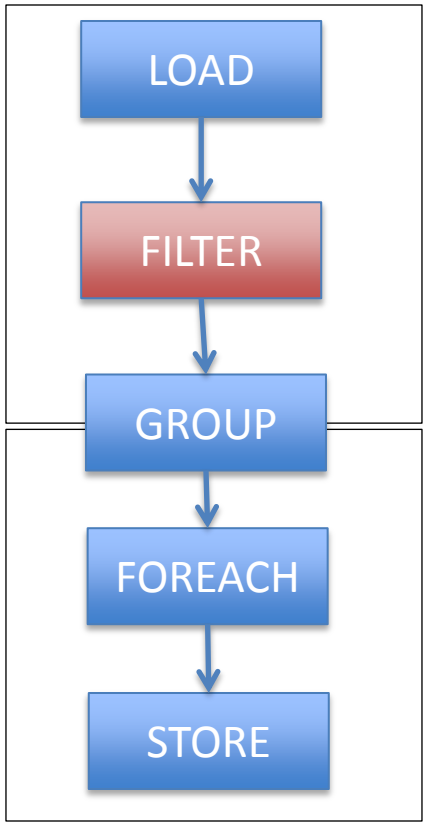
Example



Create a MR job for each COGROUP

Adapted from slides by Oliver Kennedy, U of Buffalo

Example



Map

Reduce

1) Create a MR job for each COGROUP

2) Add other commands where possible

Certain commands require their own MR job (e.g., ORDER)

Adapted from slides by Oliver Kennedy, U of Buffalo

Review

- NoSQL
  - "NoSchema", "NoTransactions", "NoLanguage"
  - A "reboot" of data systems focusing on just high-throughput reads and writes
  - But: A clear trend towards re-introducing schemas, languages, transactions at full scale
    - Google's Spanner system, for example
- Pig
  - An RA-like language layer on Hadoop
  - But not a pure relational data model
  - "Schema-on-Read" rather than "Schema-on-write"