Dremel: Interactive Analysis of Web-Scale Datasets

Google Inc
VLDB 2010

presented by
Arka Bhattacharya

some slides adapted from various Dremel presentations on the internet

The Problem: Interactive data exploration

1. Run a MapReduce to extract billions of signals from web pages

2. Ad hoc SQL against data

```
DEFINE TABLE t AS /path/to/data/*
SELECT TOP(signal, 100), COUNT(*) FROM t
. . .
```

Want answer in a few seconds (OLAP/BI).
Assumptions : Read-only, Results not too large.

…according to a Google-er

… I couldn't use it (MapReduce) when I needed nearly instantaneous results because it was too slow. Even a simple job would take several minutes to finish ….

…. simply put, if I had only used MapReduce, I couldn't have gone home until late at night …. by using Dremel I could finish by lunch time. *And if you have ever eaten lunch at Google, you know that's a big deal.*

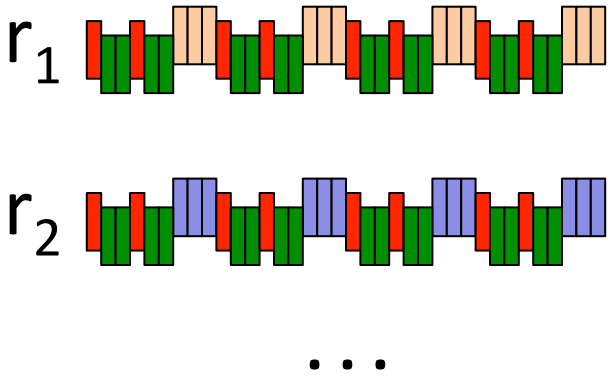
- BigQuery whitepaper

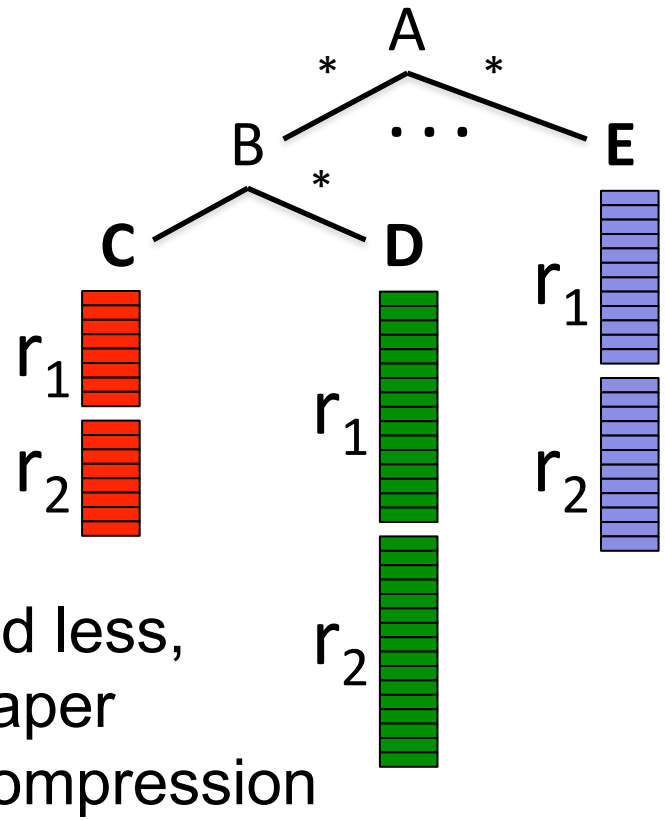
Widely used inside Google

- Analysis of crawled web documents
- Tracking install data for applications on Android Market
- Crash reporting for Google products
- OCR results from Google Books
- Spam analysis
- Debugging of map tiles on Google Maps

Idea(1) : Column-striped representation

```
DocId: 10                    r1
Links
  Forward: 20
Name
  Language
    Code: 'en-us'
    Country: 'us'
  Url: 'http://A'
Name
  Url: 'http://B'
```





Read less, cheaper decompression

Column stores for OLAP not a new idea.

Challenge: encoding nested structure of objects efficiently

Idea(1) : Column-striped representation

```
DocId: 10
Links
  Forward: 20
  Forward: 40
  Forward: 60
Name
  Language
    Code: 'en-us'
    Country: 'us'
  Language
    Code: 'en'
  Url: 'http://A'
Name
  Url: 'http://B'
Name
  Language
    Code: 'en-gb'
    Country: 'gb'
```

Links.Forward

value	r	d
20	0	2
40	1	2
60	1	2
80	0	2

Links.Backward

value	r	d
NULL	0	1
10	0	2
30	1	2

Name.Language.Code

value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

```
DocId: 20
Links
  Backward: 10
  Backward: 30
  Forward:  80
Name
  Url: 'http://C'
```

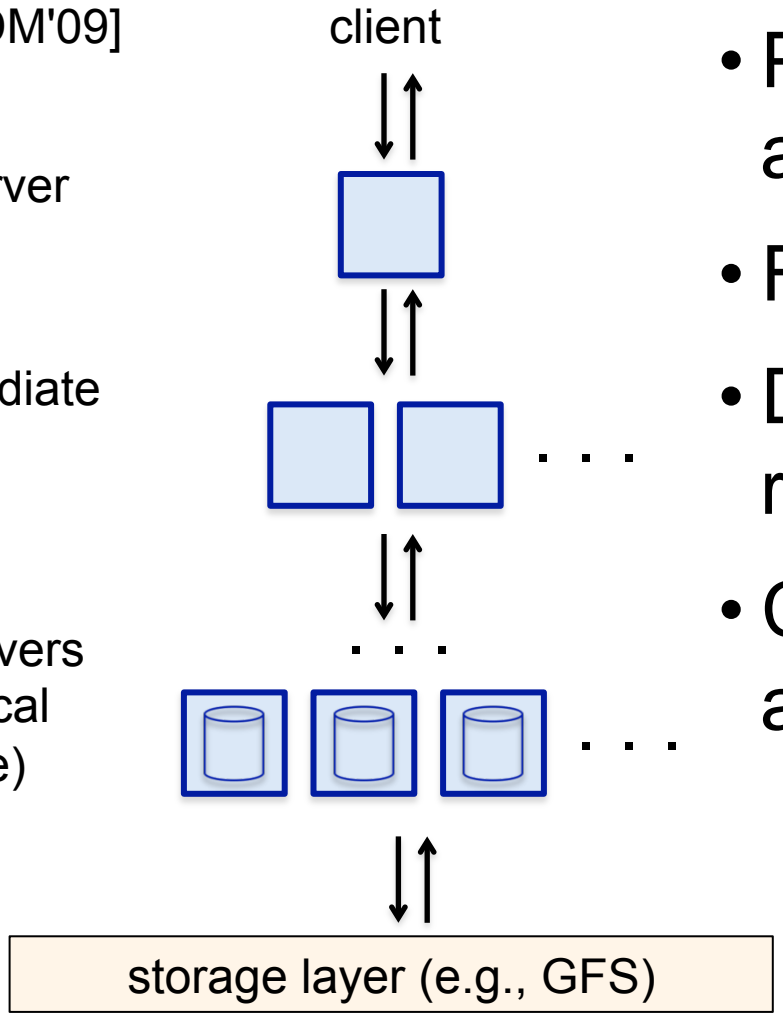
r_2

r: At what repeated field in the field's path the value has repeated

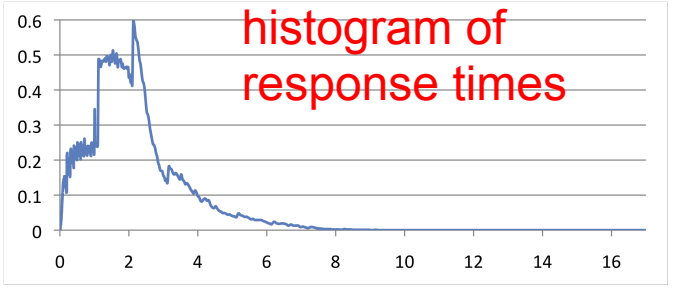
d: How many fields in paths that could be undefined (opt. or rep.) are actually present

Idea(2): Execution Tree (ala serving web-requests)

[Dean WSDM'09]

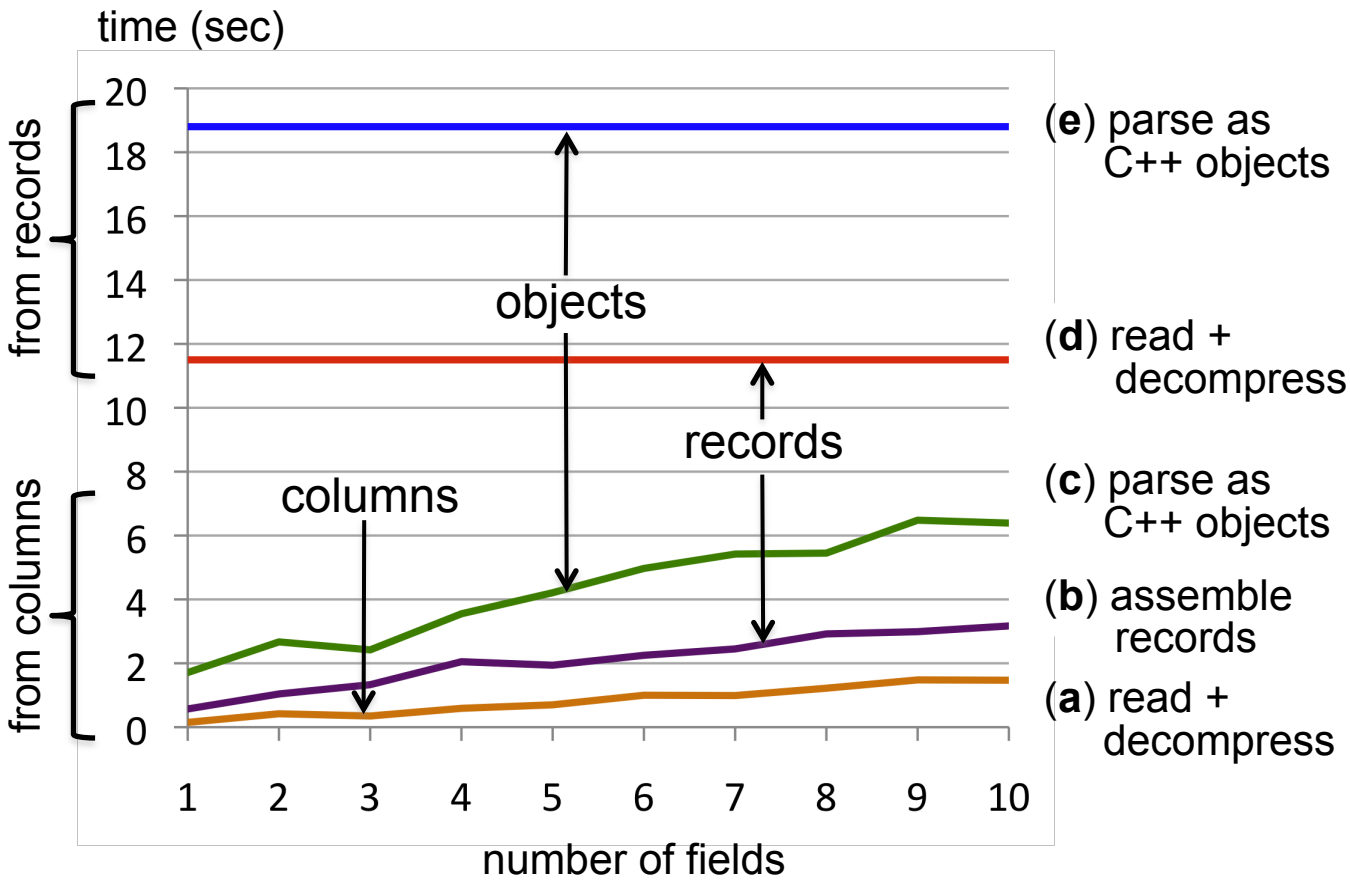


- Parallelizes scheduling and aggregation
- Fault tolerance
- Designed for "small" results (<1M records)
- Can do some approximate querying.



Read from disk

Table partition: 375 MB (compressed),
300K rows, 125 columns

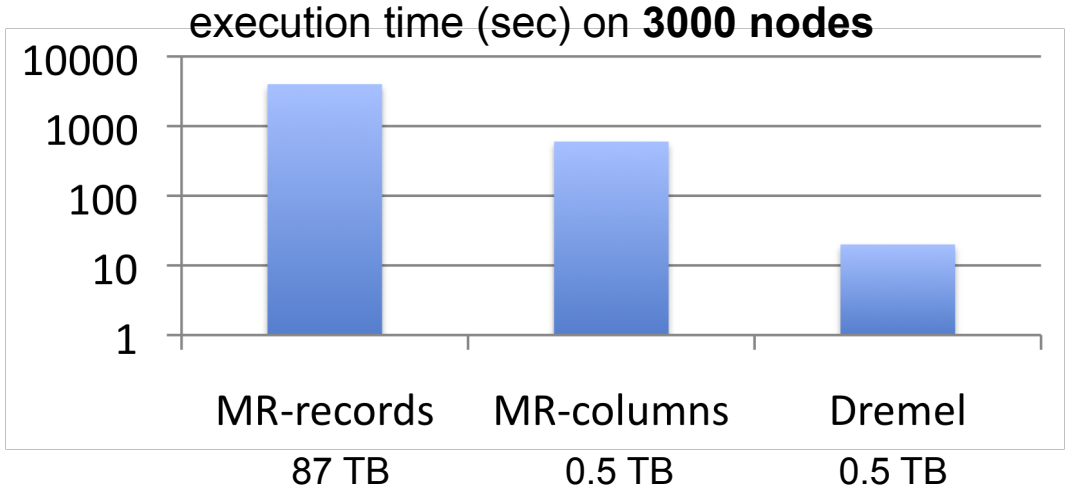


Adv of Columnar stores :

Read only Required columns + Operations on Compressed data

MR and Dremel execution

Avg # of terms in txtField in 85 billion record table T1




```
SELECT SUM(count_words(txtField)) / COUNT(*)
  FROM T1
```

State of the art at the time

- MapReduce : Processing big data vs ad-hoc interactive analysis of big data.
  - MapReduce: row-oriented, scheduling , assembling records.

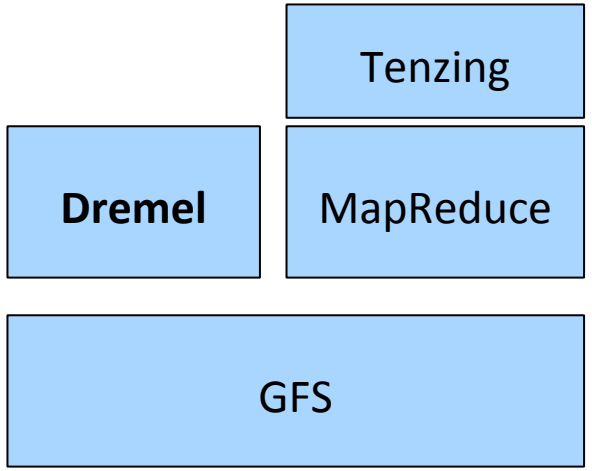
- Pig, Hive :
  - run mapreduce programs to execute query.

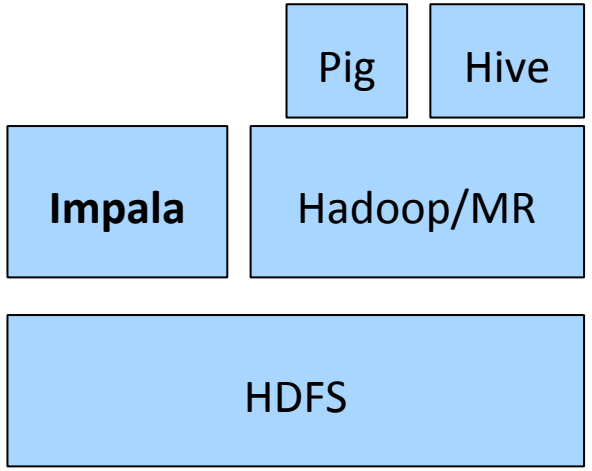
Dremel: First SQL-like query execution framework for massive datasets independent of MapReduce.

What did Dremel give up?

Dremel does a few things, but does them well !

- Updates.
  - Dremel's solution … Dont care about updates.
- Power:
  - Building a SQL implementation on top of mapreduce vs building separate in-situ query execution engine :
  - faster, but can handle (only structured) data with small result sets (e.g no large joins), and a smaller subset of SQL.
- Combined programming model
  - Unlike SparkSQL or Pig, Dremel cannot combine procedural programming with SQL-like declarative programming.
- Global query optimization ?
  - Not a lot of query cost optimization details provided in the paper.





Impact!

- In-use at Google since 2006 !
- Apache Drill:
  - Open source implementation of Dremel.
- BigQuery
  - Commercial offering by Google with Dremel underneath.
- Nested columnar storage inspired columnar file formats such as Parquet.