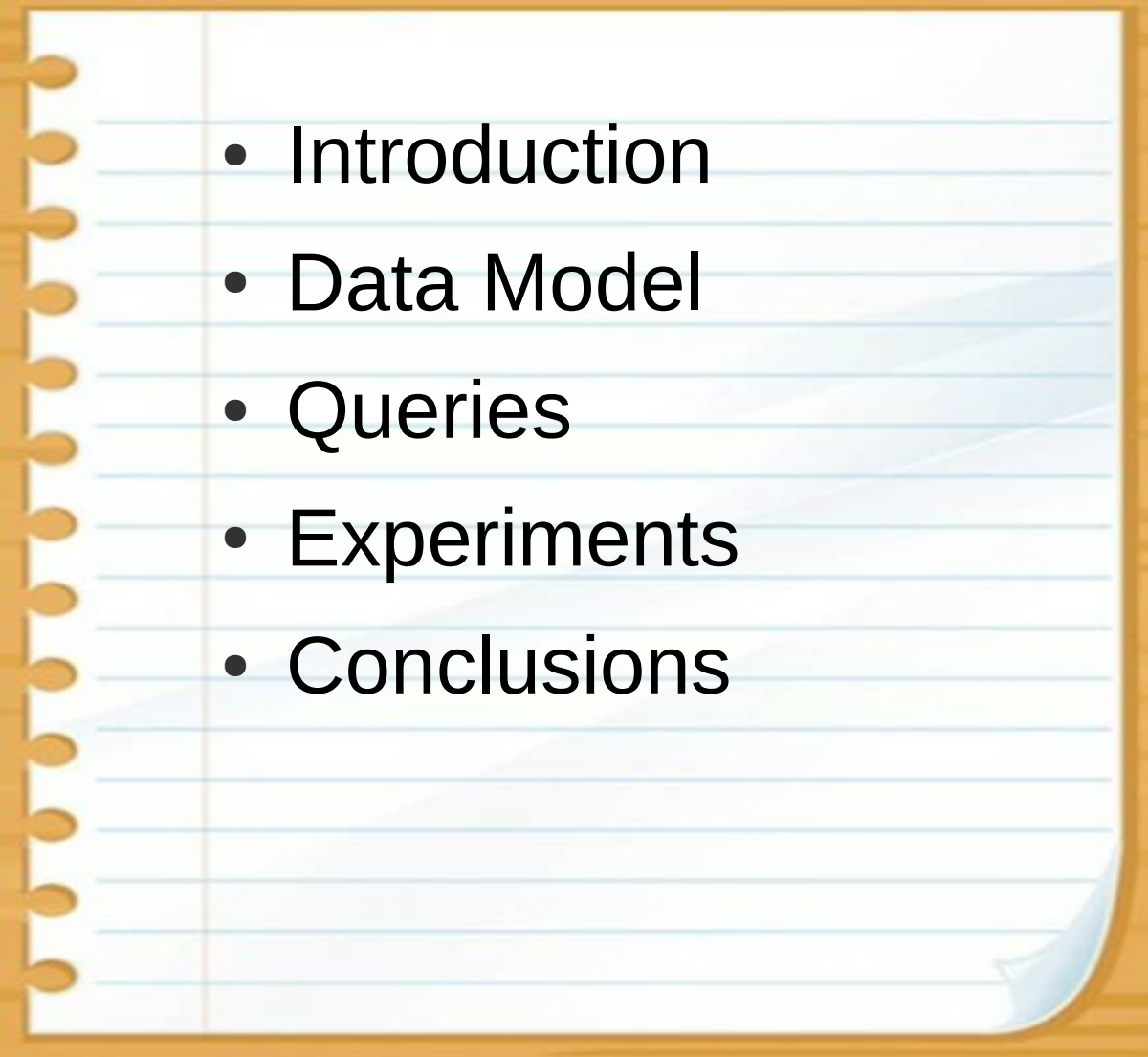


Dremel

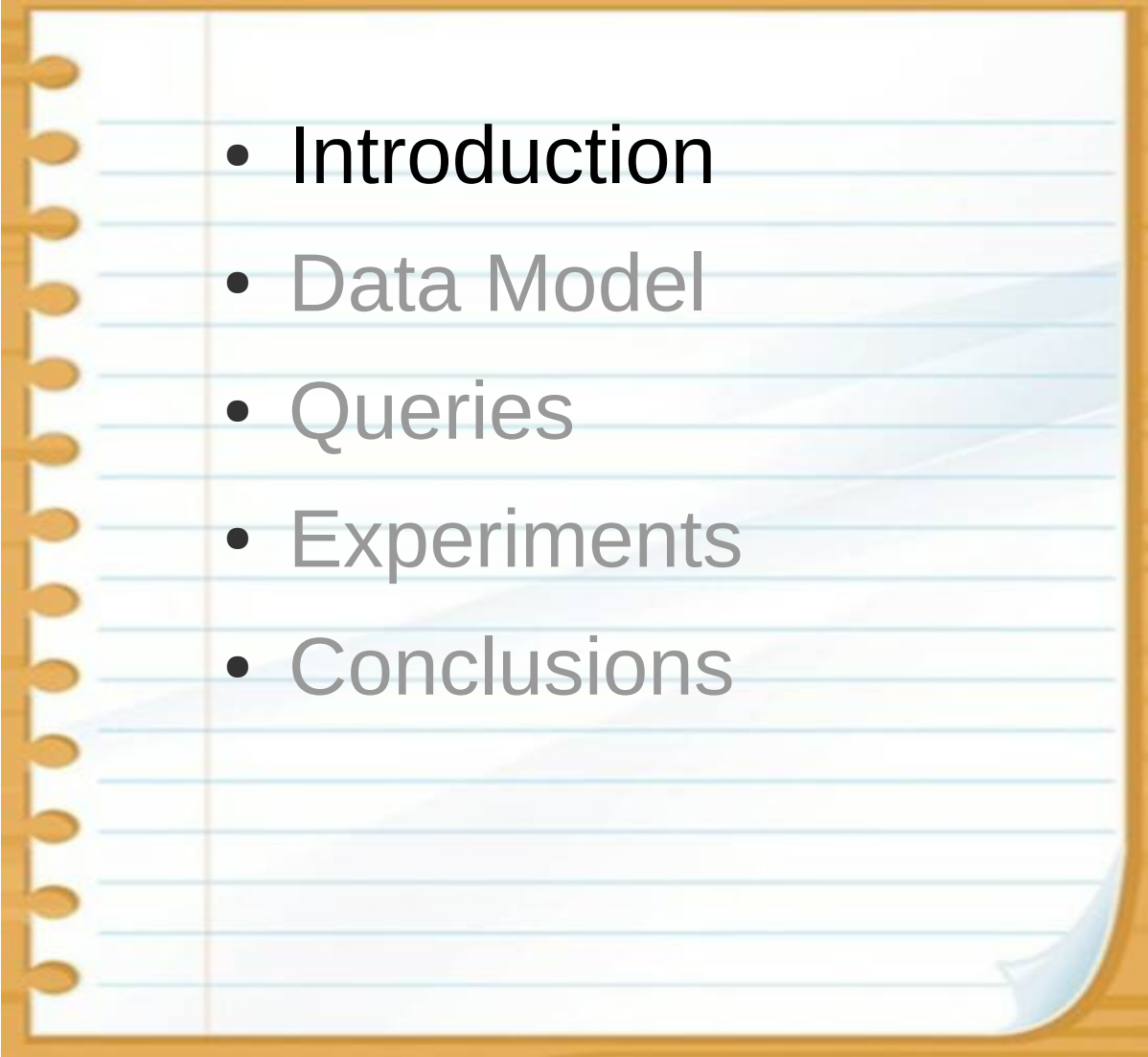
Interactive Analysis of Web-Scale Datasets

Vasia Kalavri (kalavri@kth.se)
EMDC 2011 – Advanced DS

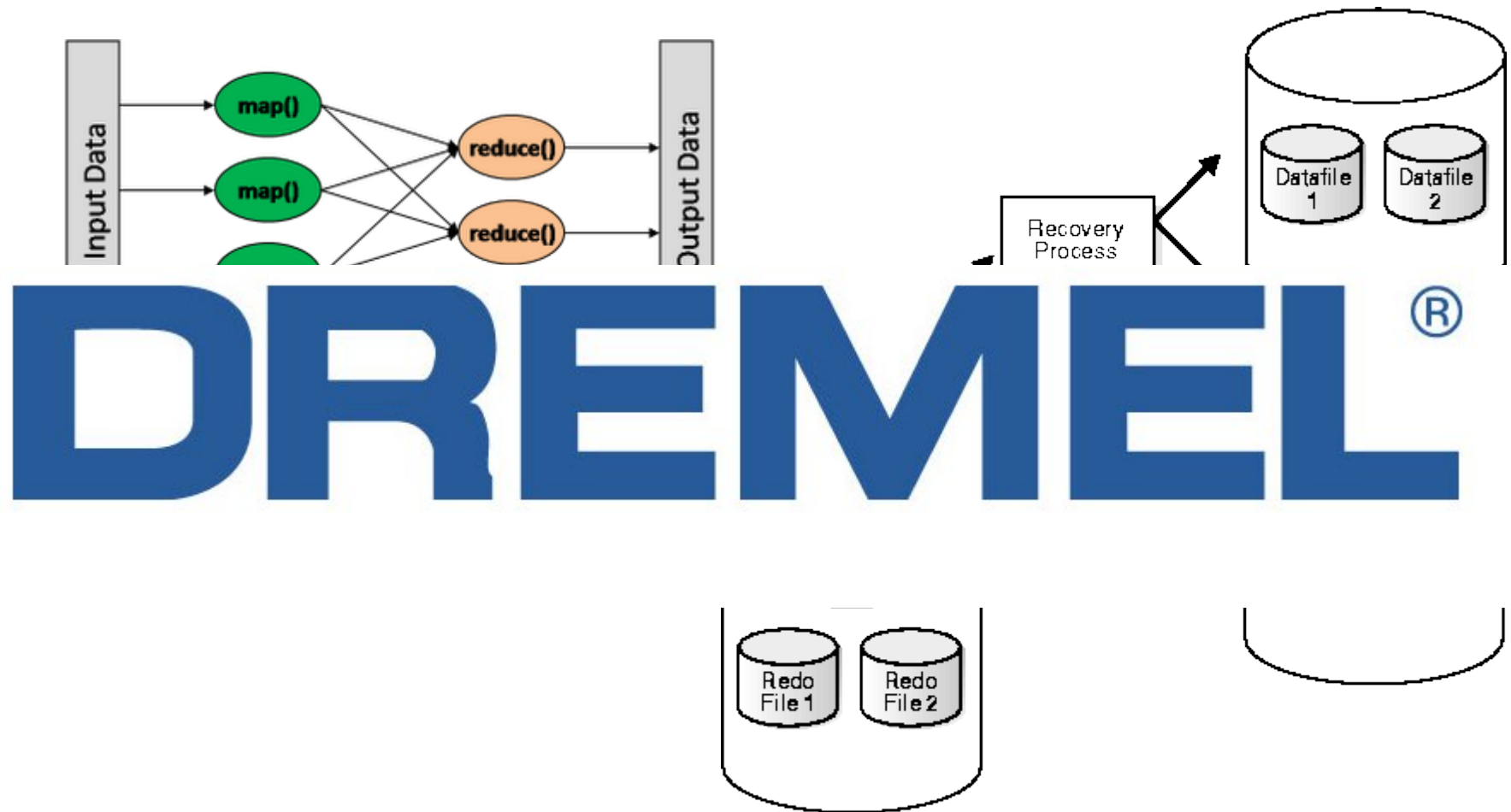
Outline

- 
- Introduction
 - Data Model
 - Queries
 - Experiments
 - Conclusions

Outline

- 
- A spiral-bound notebook with a light blue cover and a white page with horizontal blue lines. The notebook is open, and the list of topics is written on the right page. The spiral binding is on the left side.
- Introduction
 - Data Model
 - Queries
 - Experiments
 - Conclusions

Yet Another BigData Analytics Tool?



System Goals

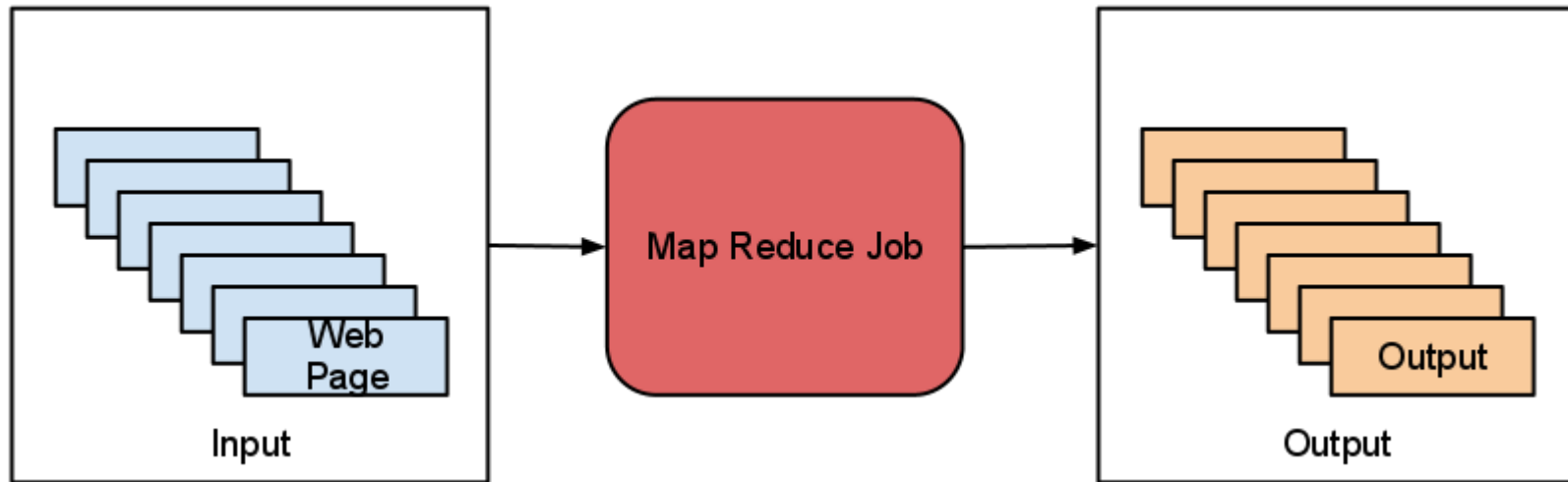
- Scalability
- Interactivity
- Aggregation Queries
- NOT a MR replacement but a complement

Use Case



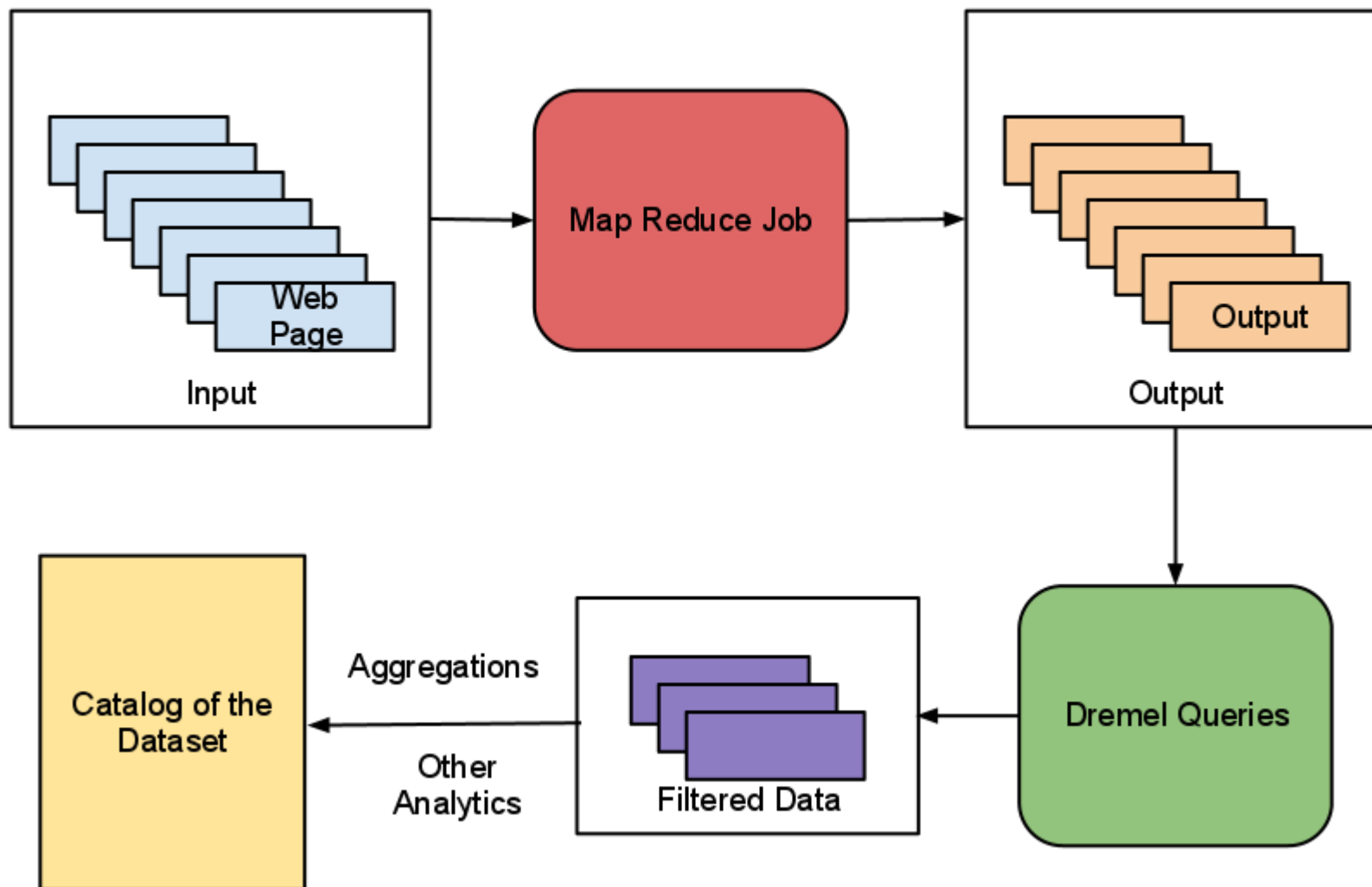
Extract new kinds
of signals from
Web Pages!

Use Case

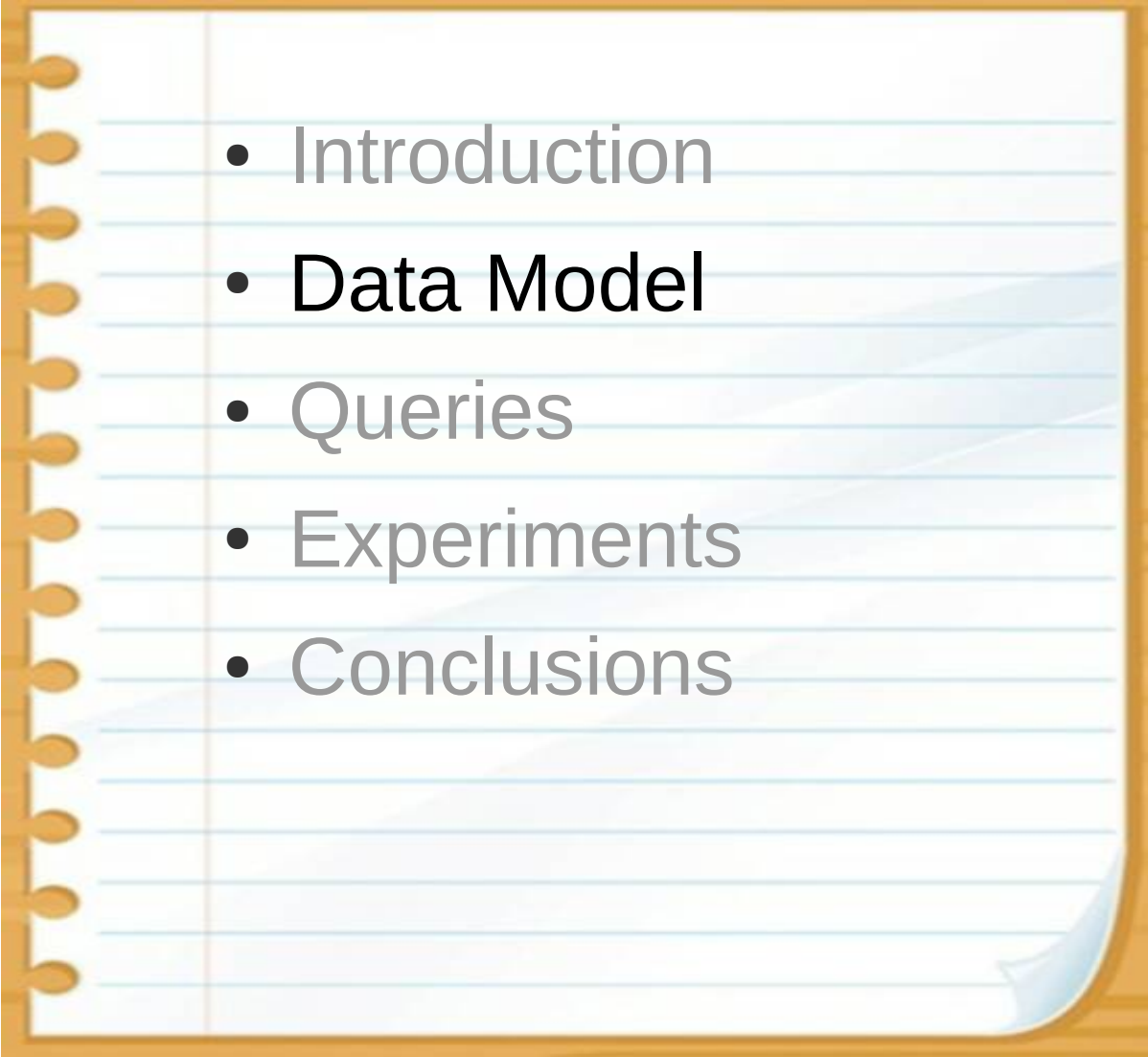


- The output is distributed in the file system..
- How do I extract the signals?

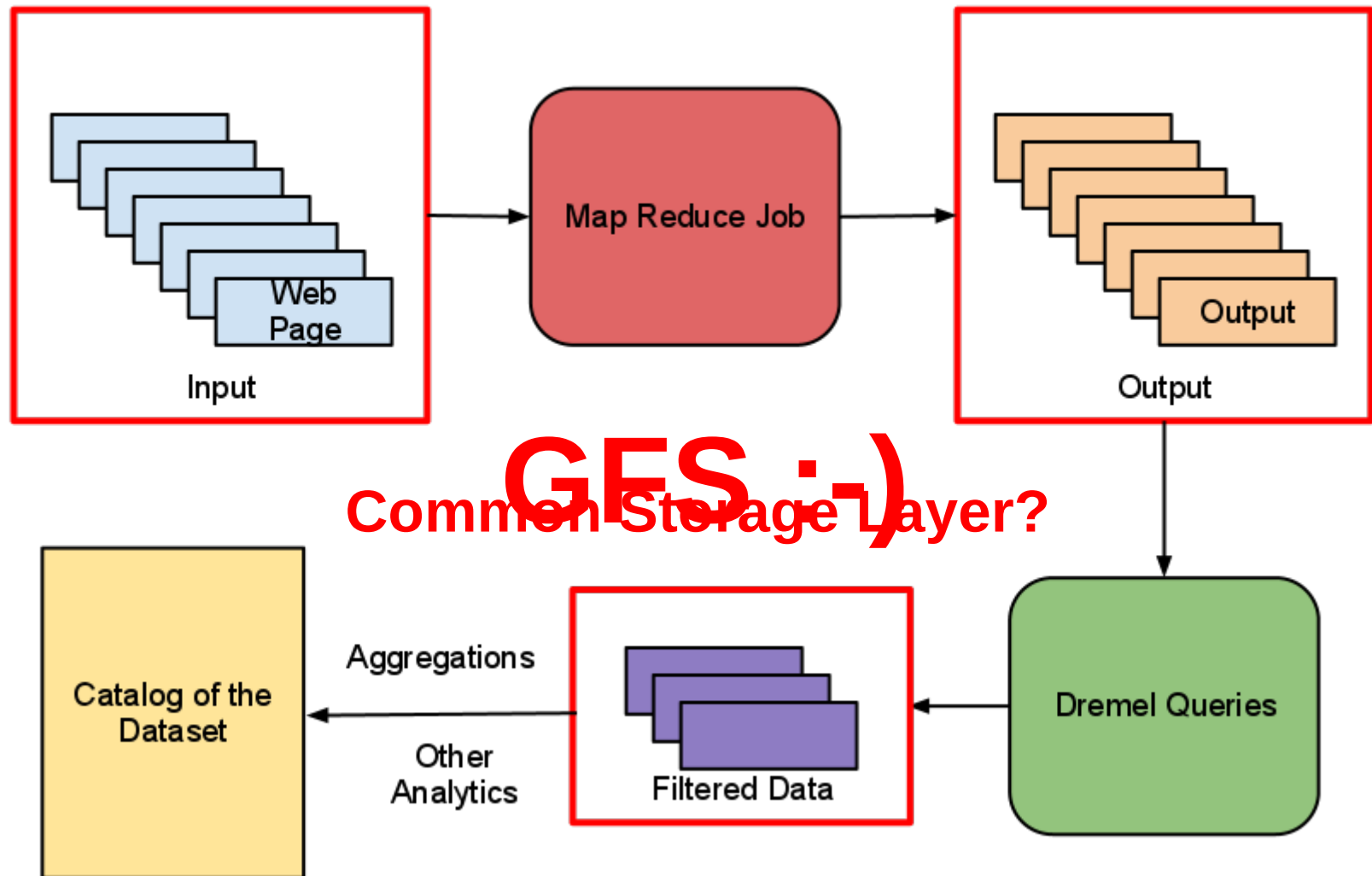
Use Case



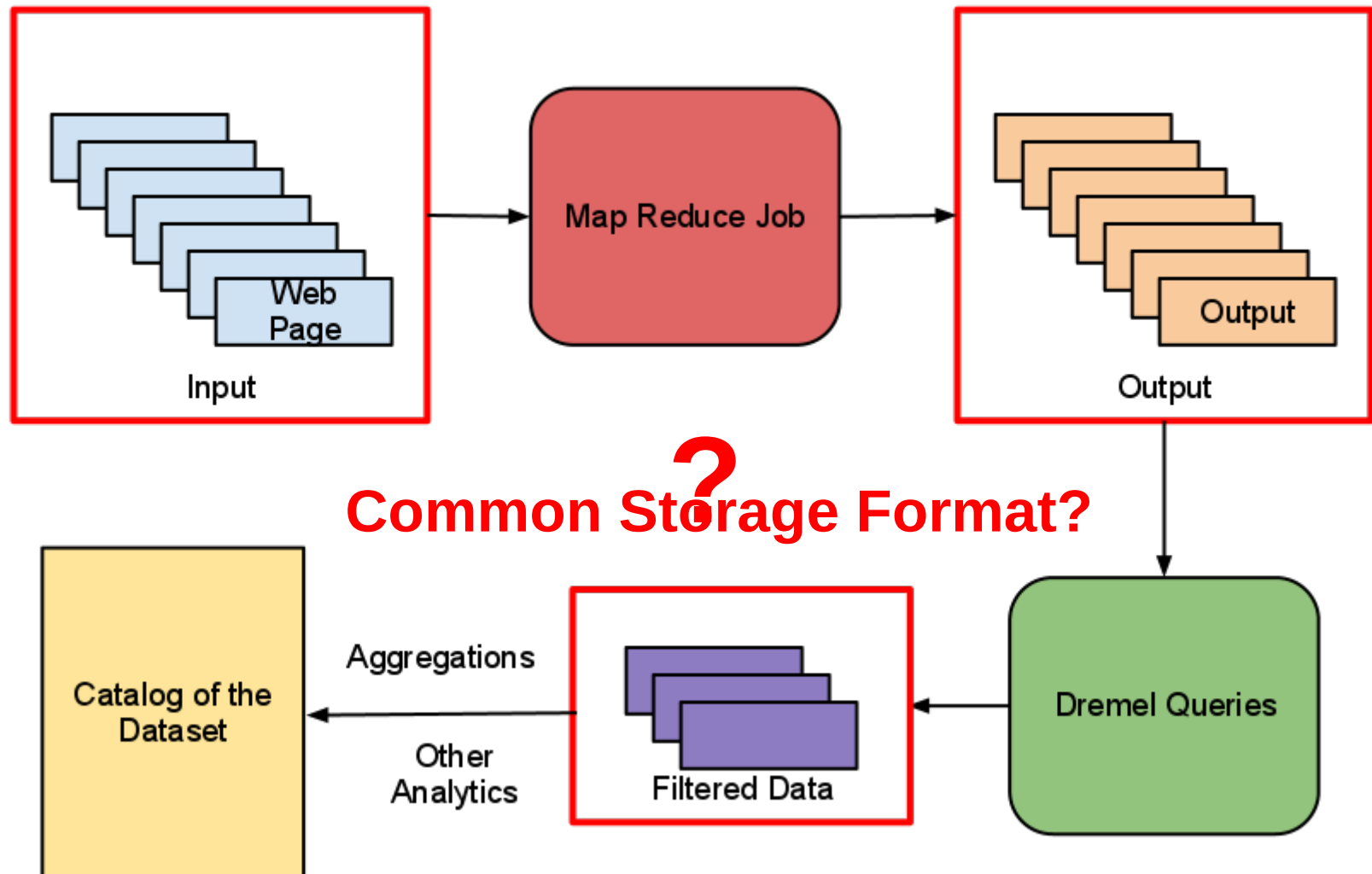
Outline

- 
- Introduction
 - **Data Model**
 - Queries
 - Experiments
 - Conclusions

Use Case - revisited



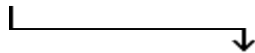
Use Case - revisited



Data Data Data

Once Upon a Time...

Id	Name	Surname
300	Vasia	Kalavri



Id	F_Id	Department
321	300	IT

Records of **Relational** Data

But.. Web Data

Website: "http://www.lol.com"

Page

Name: "index"

URL: "index.html"

Link

URL: "/a.com"

Link

URL: "/b.com"

Page

Name: "contact"

URL: "contact.html"

Link

URL: "c.com"

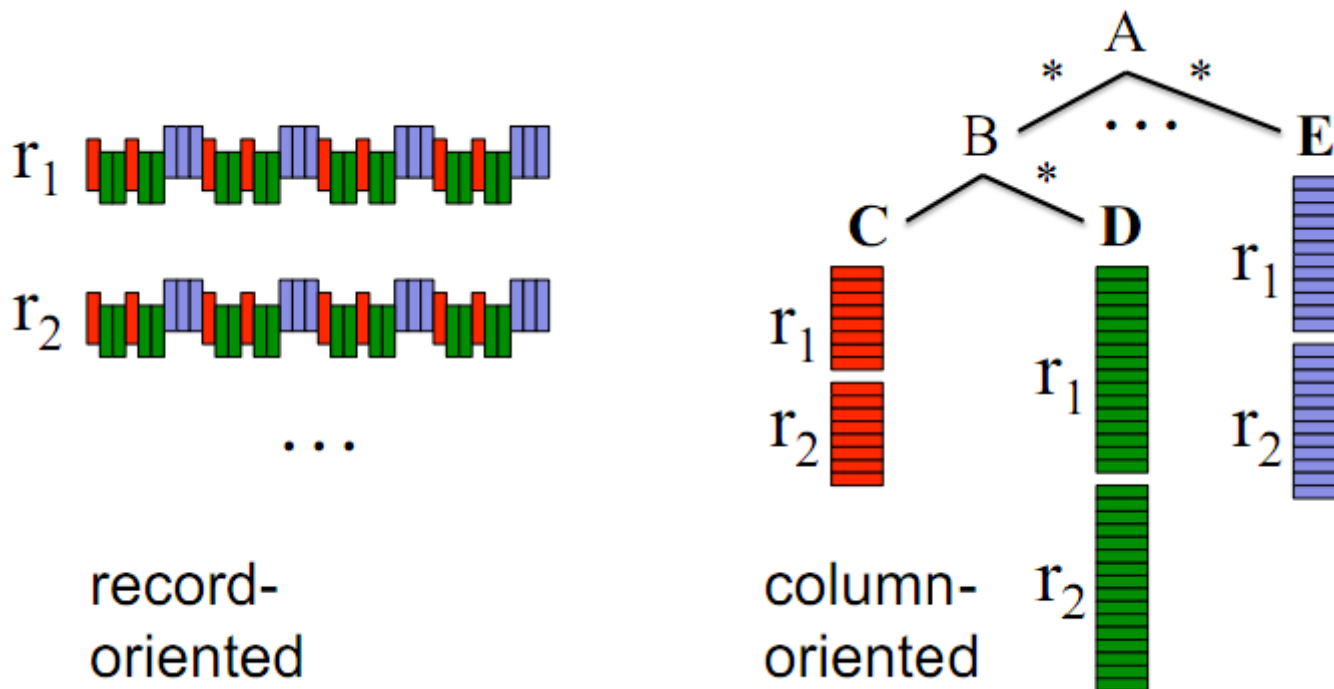
Link

Email: "info@boo.com"

is Nested!

Columnar Storage

- Successful for flat Relational Data
- Adapt for Nested Data?



Columnar Storage - Challenges

- How to...
 - Represent data in a **lossless** way?
 - Encode data **fast**?
 - Assemble records **efficiently**?

Lossless Representation (1)

- Field Types:
 - Required
 - Repeated
 - Optional
- We attach 2 *levels* to each value:
 - Repetition Level
 - Definition Level

Lossless Representation (2)

```
message Document {  
  required int64 DocId;  
  optional group Links {  
    repeated int64 Backward;  
    repeated int64 Forward; }  
  repeated group Name {  
    repeated group Language {  
      required string Code;  
      optional string Country; }  
    optional string Url; } }
```

DocId: 10 **r₁**
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'

Name.Language.Code		
value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

Record Encoding

- Recursive Algorithm
 - Traverses the record
 - Computes Levels for each value

```
DocId: 10      r1
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'
```



Name.Language.Code		
value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

Record Assembly

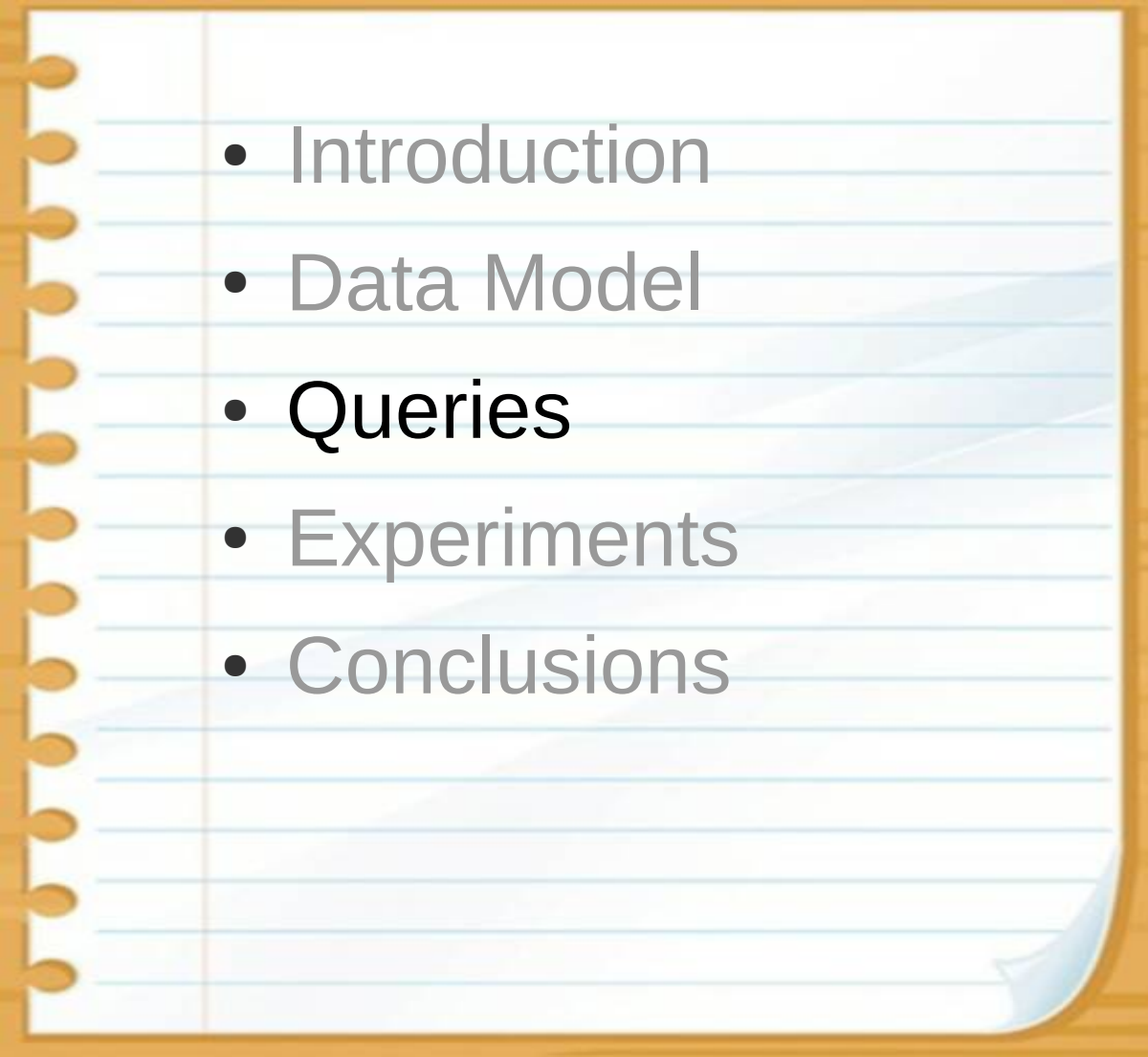
- FSM
 - Reads values and levels
 - Appends values sequentially to the output records

```
DocId: 10      r1
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'
```



Name.Language.Code		
value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

Outline

- 
- A spiral-bound notebook with a light blue cover and a white page with horizontal blue lines. The notebook is open, and the list of topics is written on the right page. The topics are: Introduction, Data Model, Queries, Experiments, and Conclusions. The word 'Queries' is bolded.
- Introduction
 - Data Model
 - **Queries**
 - Experiments
 - Conclusions

Query Language

- Based on SQL
- Refined for columnar nested storage

```
SELECT DocId AS Id,  
       COUNT(Name.Language.Code) WITHIN Name AS Cnt,  
       Name.Url + ',' + Name.Language.Code AS Str  
FROM t  
WHERE REGEXP(Name.Url, '^http') AND DocId < 20;
```


Query Execution (1)

- Tree architecture
 - Root Server
 - Receives incoming queries
 - Reads table metadata
 - Routes queries to the next level of the tree
 - Leaf Servers
 - Communicate with storage layer

Query Execution (2)

- Definitions
 - **Slot:** *Unit available for execution*
 - **Tablet:** *Horizontal partition of a table*
- Query Dispatcher
 - Schedules queries to slots
 - Balances the load
 - Assures fault-tolerance
 - Specifies what percentage of tablets need to be scanned before returning a result

Outline

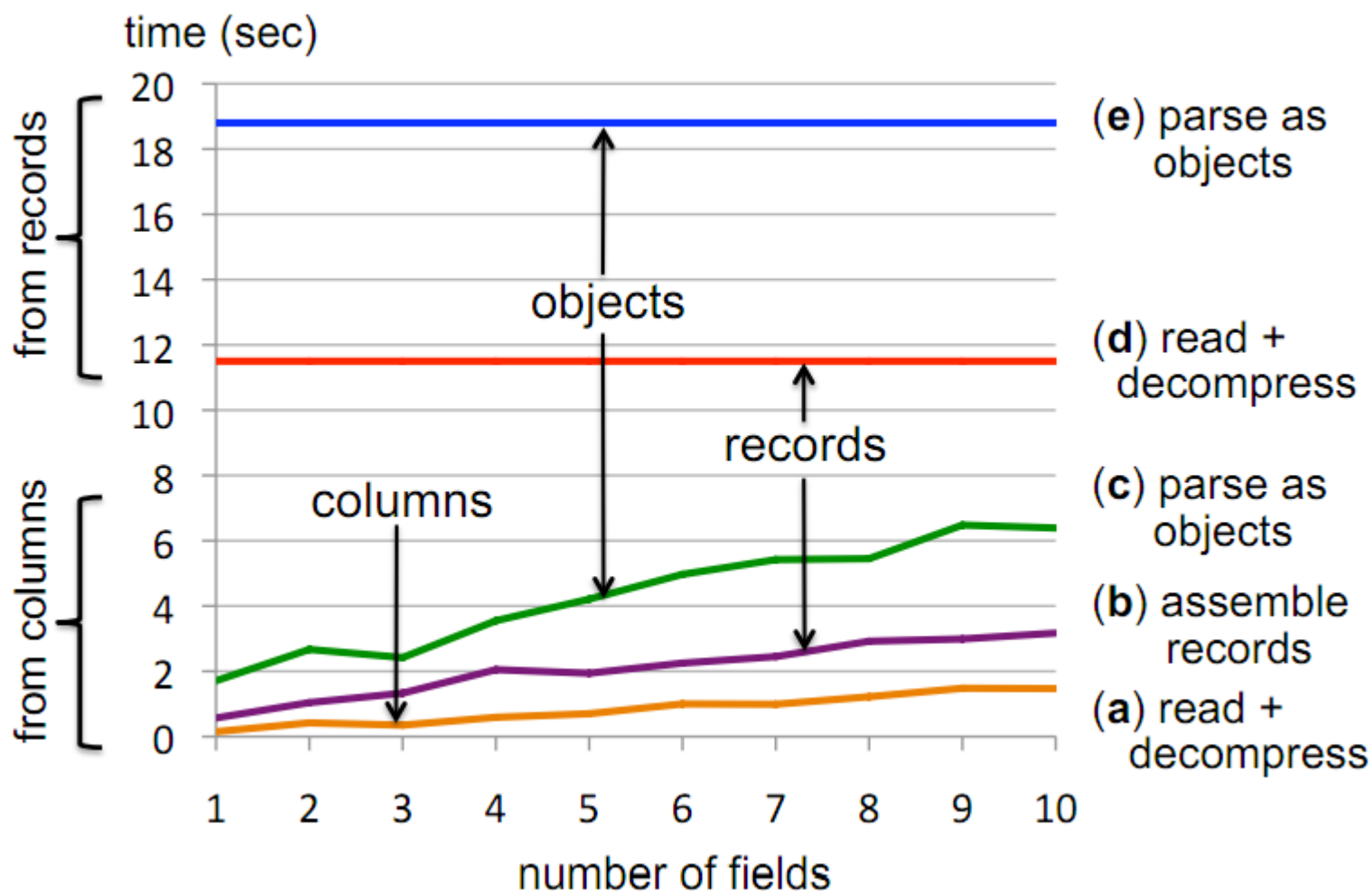
- 
- A spiral-bound notebook with a light blue cover and a white page with horizontal blue lines. The notebook is open, and the page is showing a list of topics. The topics are: Introduction, Data Model, Queries, Experiments, and Conclusions. The word 'Experiments' is bolded. The notebook has a gold-colored spiral binding on the left side.
- Introduction
 - Data Model
 - Queries
 - **Experiments**
 - Conclusions

Experiments

- Experiment Types
 - Data access characteristics
 - Columnar Storage for MR
 - Dremel's performance
- *Instances running on two data centers during regular business operation*

Local disk

- Columnar vs. record-oriented storage



MR & Dremel

- Computes an average
- 0,5TB by MR-on-columns
- 87TB by MR-on-records

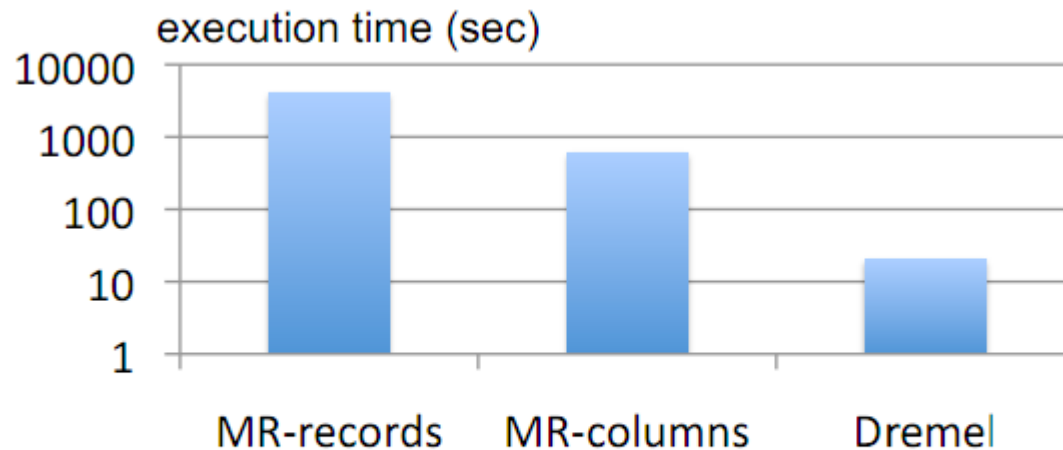


Figure 10: MR and Dremel execution on columnar vs. record-oriented storage (3000 nodes, 85 billion records)

Aggregation Trees

- 24 billion nested records
- Q2: Read-intensive aggregation
- Q3: Write-intensive aggregation

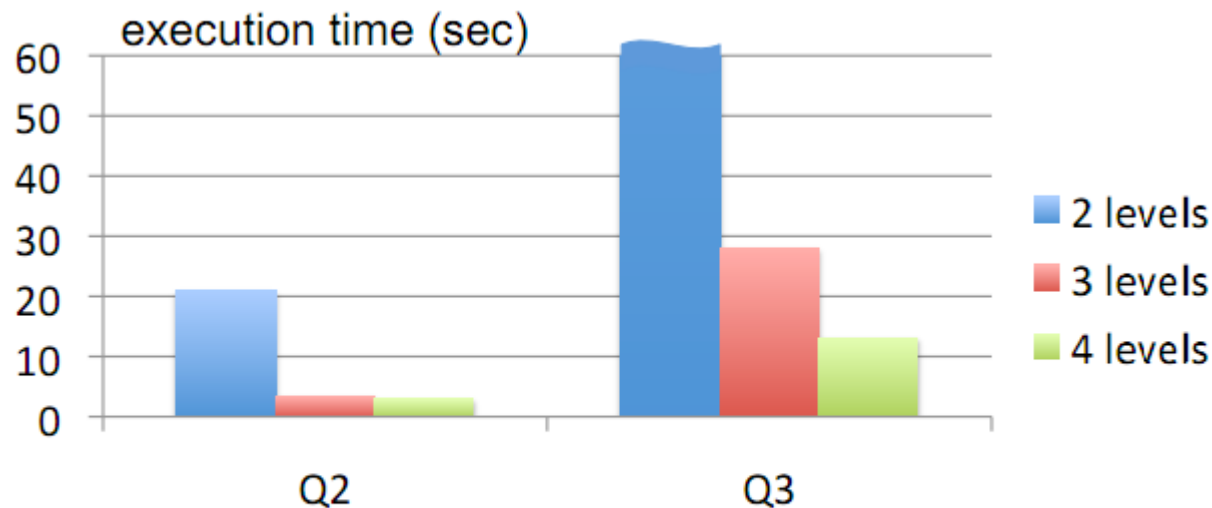
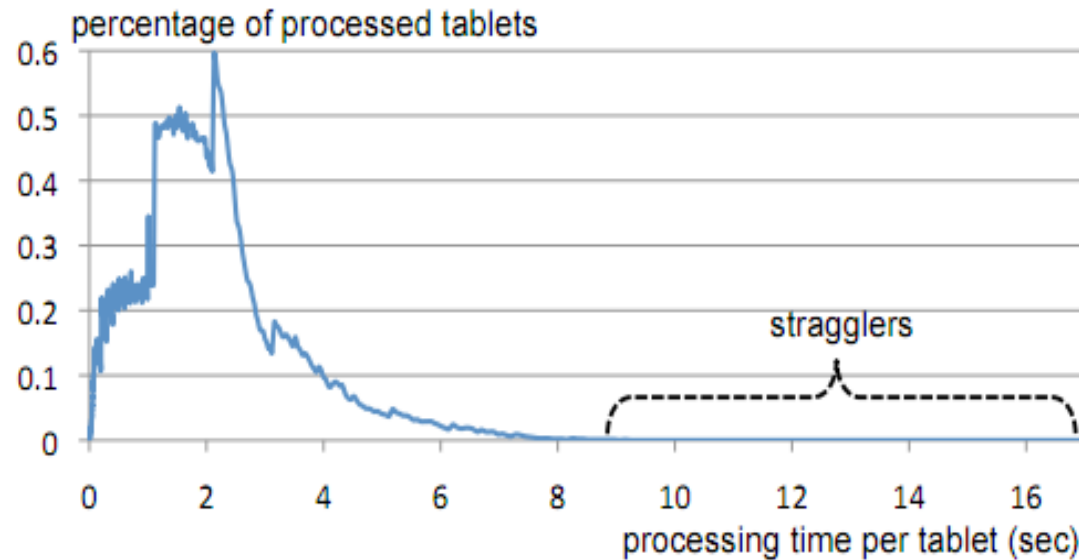



Figure 11: Execution time as a function of serving tree levels for two aggregation queries on T_2

Speed vs. Accuracy

- 99% of the results are computed within 5s



Outline

- 
- Introduction
 - Data Model
 - Queries
 - Experiments
 - **Conclusions**

Conclusions (1)

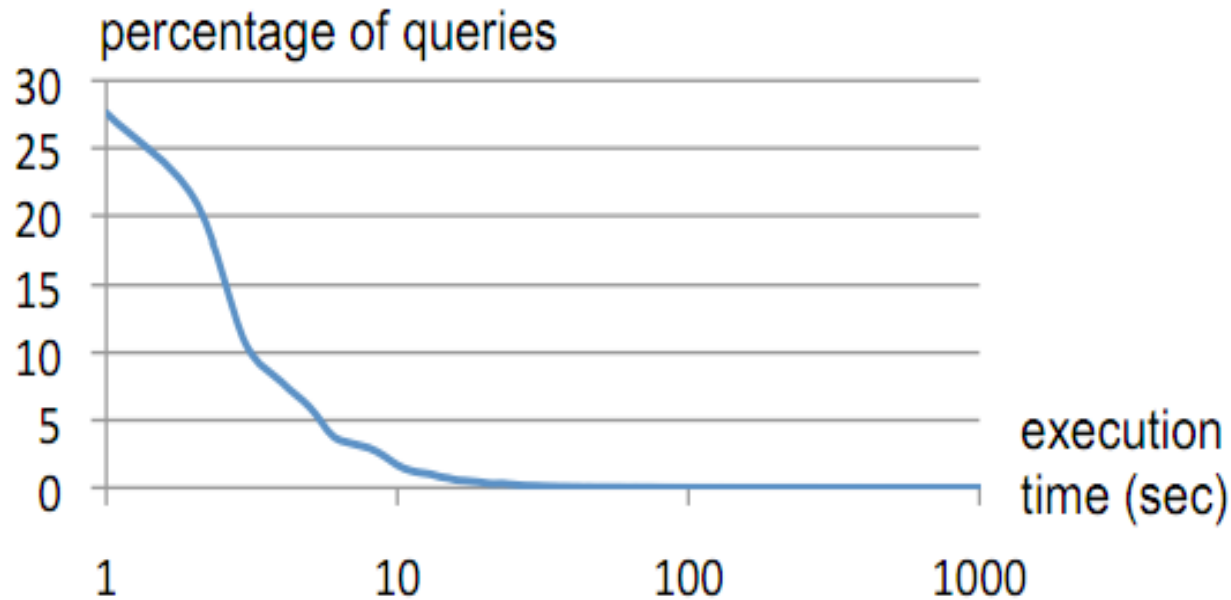


Figure 15: Query response time distribution in a monthly workload

Conclusions (2)

- On Columnar Storage
 - Good fit for nested data
 - Even MR can benefit from it
- On Interactivity
 - Dremel is interactive when running aggregation queries on a small amount of fields
 - Trade speed against accuracy!
- Dremel and MR can be effectively used in a complementary fashion

Dremel

Interactive Analysis of Web-Scale Datasets

Vasia Kalavri (kalavri@kth.se)
EMDC 2011 – Advanced DS