

BigQuery explained

The Dremel white paper for dummies



Dirk Primbs
Google, Developer Relations

BigQuery Demo

[->link](#)

using public GitHub timeline data

The screenshot shows the Google BigQuery web interface. At the top right, the text "cloud_dataengine_20131018_0_RC00" is visible. The main header features the "Google bigquery" logo. On the left sidebar, there are buttons for "COMPOSE QUERY", "Query History", and "Job History". Below these, a section titled "BigQuery demo" shows a dropdown menu with "githubarchive:github" and "publicdata:samples" options. The main area displays a SQL query titled "GitHub commit swearing" with a "? X" icon in the top right corner. The query is as follows:

```
1 SELECT
2   repository_language,
3   COUNT(1) AS swearing_commits
4 FROM
5   [githubarchive:github.timeline]
6 WHERE
7   repository_language IS NOT NULL
8   AND payload_commit_msg IS NOT NULL
9   AND REGEXP_MATCH (payload_commit_msg,
10    r'(?i)\b(fuck|fucked|fucking|damn|damned|shit|shitty|beastard|hell|clusterfuck
11 GROUP BY
12   repository_language
13 ORDER BY
14   swearing_commits DESC
```

Below the query editor, there are four buttons: "RUN QUERY" (highlighted in red), "Save Query", "Prettify Query", and "Enable Options". A status message at the bottom reads "Query complete (12.9s elapsed, 4.08 GB processed)" and a green checkmark icon is in the bottom right corner.

BOOK 1

AUTHOR: Dumas

TITLE: Three Musketeers

PRICE

DISCOUNT: 0

USD: 20

EUR: 19

BOOK 2

AUTHOR: Yrsa Sigurdardottir

AUTHOR: Tina Flecken

AUTHOR: Elma Klein

TITLE: Feuernacht

BOOK 3

TITLE: Stay Fit

PRICE

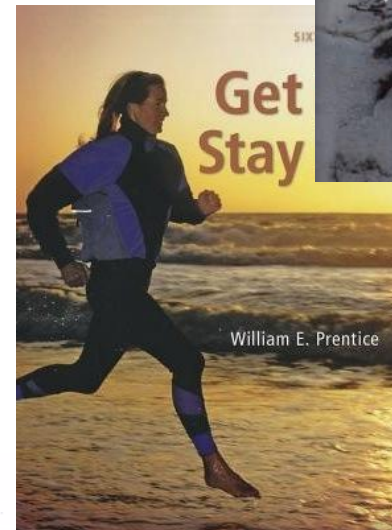
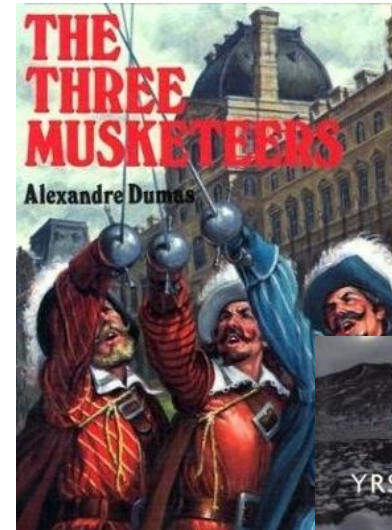
DISCOUNT: 0

EUR: 12

PRICE

DISCOUNT: 1

EUR: 11



```
SELECT Count(*) FROM Books  
WHERE (PRICE.EUR > 9 OR PRICE.USD > 10)
```

Columnar representation

AUTHOR

Dumas (0, 1)
Yrsa Sigurdardottir (0, 1)
Tina Flecken (1, 1)
Elma Klein (1, 1)
NULL (0, 0)

PRICE.EUR

19 (0, 2)
NULL (0, 0)
12 (0, 2)
11 (1, 2)

PRICE.DISCOUNT

0 (0, 2)
NULL (0, 0)
0 (0, 2)
1 (1, 2)

TITLE

Three Musketeers (0,1)
Feuernacht (0, 1)
Stay Fit (0, 1)

PRICE.USD

20 (0, 2)
NULL (0, 0)
NULL (0, 1)
NULL (1, 1)



BOOK 1

AUTHOR: Dumas

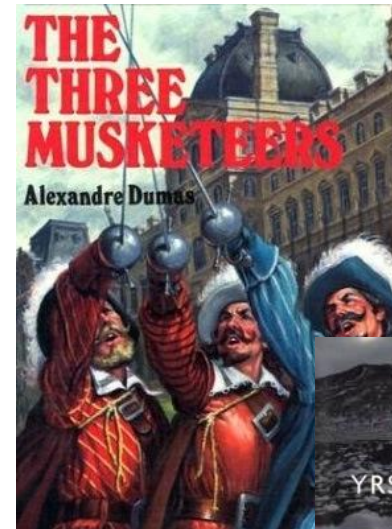
TITLE: Three Musketeers

PRICE

DISCOUNT: 0

USD: 20

EUR: 19



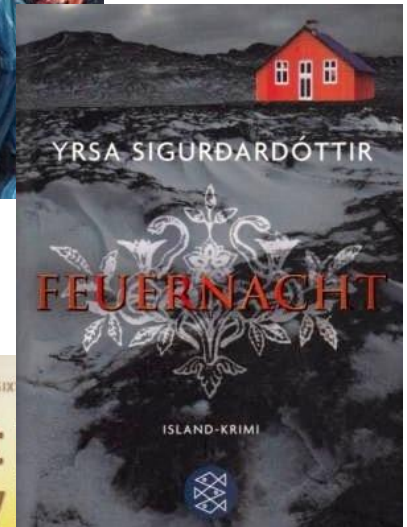
BOOK 2

AUTHOR: Yrsa Sigurdardottir

AUTHOR: Tina Flecken

AUTHOR: Elma Klein

TITLE: Feuernacht



BOOK 3

TITLE: Stay Fit

PRICE

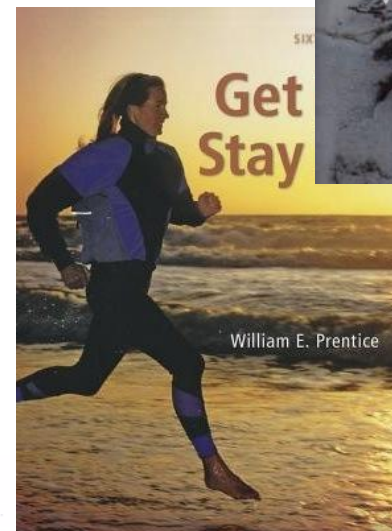
DISCOUNT: 0

EUR: 12

PRICE

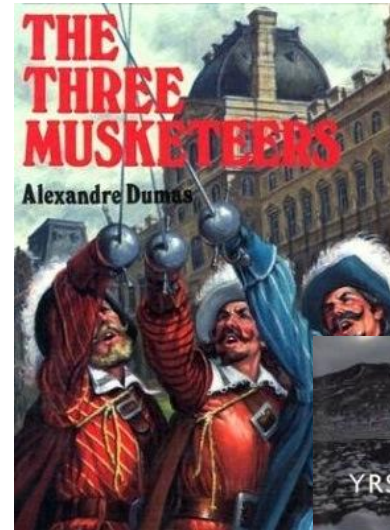
DISCOUNT: 1

EUR: 11



BOOK 1

AUTHOR: Dumas
TITLE: Three Musketeers
PRICE
DISCOUNT: 0
USD: 20
EUR: 19



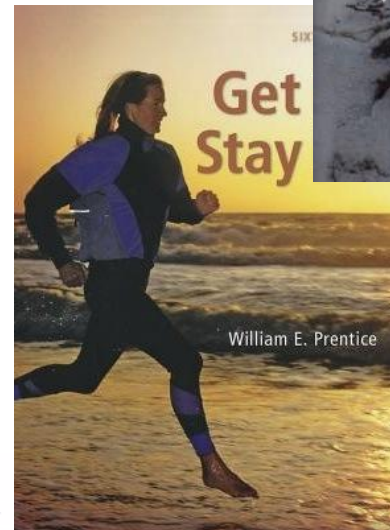
BOOK 2

AUTHOR: Yrsa Sigurdardottir
AUTHOR: Tina Flecken
AUTHOR: Elma Klein
TITLE: Feuernacht
(PRICE)
(DISCOUNT): NULL
(EUR): NULL
(USD): NULL



BOOK 3

(AUTHOR): NULL
TITLE: Stay Fit
PRICE
DISCOUNT: 0
(USD): NULL
EUR: 12
PRICE
DISCOUNT: 1
(USD): NULL
EUR: 11



BOOK 1

AUTHOR:	Dumas	AUTHOR
TITLE:	Three Musketeers	TITLE
PRICE		
DISCOUNT:	0	PRICE.DISCOUNT
USD:	20	PRICE.USD
EUR:	19	PRICE.EUR

BOOK 2

AUTHOR:	Yrsa Sigurdardottir	AUTHOR
AUTHOR:	Tina Flecken	AUTHOR[1]
AUTHOR:	Elma Klein	AUTHOR[2]
TITLE:	Feuernacht	TITLE
(PRICE)		
(DISCOUNT):	NULL	(PRICE).(DISCOUNT)
(EUR):	NULL	(PRICE).(EUR)
(USD):	NULL	(PRICE).(USD)

BOOK 3

(AUTHOR):	NULL	(AUTHOR)
TITLE:	Stay Fit	TITLE
PRICE		
DISCOUNT:	0	PRICE.DISCOUNT
(USD):	NULL	PRICE.(USD)
EUR:	12	PRICE.EUR
PRICE		
DISCOUNT:	1	PRICE[1].DISCOUNT
(USD):	NULL	PRICE[1].(USD)
EUR:	11	PRICE[1].EUR

BOOK 1

AUTHOR: Dumas
 TITLE: Three Musketeers
 PRICE
 DISCOUNT: 0
 USD: 20
 EUR: 19

AUTHOR
 TITLE
 PRICE.DISCOUNT
 PRICE.USD
 PRICE.EUR

	R	D
AUTHOR	0	1
TITLE	0	1
PRICE.DISCOUNT	0	2
PRICE.USD	0	2
PRICE.EUR	0	2

BOOK 2

AUTHOR: Yrsa Sigurdardottir
 AUTHOR: Tina Flecken
 AUTHOR: Elma Klein
 TITLE: Feuernacht
 (PRICE)
 (DISCOUNT): NULL
 (EUR): NULL
 (USD): NULL

AUTHOR
 AUTHOR[1]
 AUTHOR[2]
 TITLE
 (PRICE).(DISCOUNT)
 (PRICE).(EUR)
 (PRICE).(USD)

AUTHOR	0	1
AUTHOR[1]	1	1
AUTHOR[2]	1	1
TITLE	0	1
(PRICE).(DISCOUNT)	0	0
(PRICE).(EUR)	0	0
(PRICE).(USD)	0	0

BOOK 3

(AUTHOR): NULL
 TITLE: Stay Fit
 PRICE
 DISCOUNT: 0
 (USD): NULL
 EUR: 12
 PRICE
 DISCOUNT: 1
 (USD): NULL
 EUR: 11

(AUTHOR)
 TITLE
 PRICE.DISCOUNT
 PRICE.(USD)
 PRICE.EUR
 PRICE[1].DISCOUNT
 PRICE[1].(USD)
 PRICE[1].EUR

(AUTHOR)	0	0
TITLE	0	1
PRICE.DISCOUNT	0	2
PRICE.(USD)	0	1
PRICE.EUR	0	2
PRICE[1].DISCOUNT	1	2
PRICE[1].(USD)	1	1
PRICE[1].EUR	1	2

R = In the path to the field, what is the last repeated field ?

D = In the path to the field, how many defined fields ?

Lockstep column traversal

example on one column

AUTHOR

→Dumas (0, 1)
Yrsa Sigurdardottir (0, 1)
Tina Flecken (1, 1)
Elma Klein (1, 1)
NULL (0, 0)

BOOK 1

AUTHOR: Dumas

Lockstep column traversal

example on one column

AUTHOR

Dumas (0, 1)
→ Yrsa Sigurdardottir (0, 1)
Tina Flecken (1, 1)
Elma Klein (1, 1)
NULL (0, 0)

BOOK 1

AUTHOR: Dumas

BOOK 2

AUTHOR: Yrsa Sigurdardottir

R = 0

We have a new record

Lockstep column traversal

example on one column

AUTHOR

Dumas (0, 1)
Yrsa Sigurdardottir (0, 1)
→ Tina Flecken (1, 1)
Elma Klein (1, 1)
NULL (0, 0)

BOOK 1

AUTHOR: Dumas

BOOK 2

AUTHOR: Yrsa Sigurdardottir

AUTHOR: Tina Flecken

$R = 1$

The 1st element in the path
(i.e. AUTHOR) is repeated

Lockstep column traversal

example on one column

AUTHOR

Dumas (0, 1)
Yrsa Sigurdardottir (0, 1)
Tina Flecken (1, 1)
→ Elma Klein (1, 1)
NULL (0, 0)

BOOK 1

AUTHOR: Dumas

BOOK 2

AUTHOR: Yrsa Sigurdardottir

AUTHOR: Tina Flecken

AUTHOR: Elma Klein

$R = 1$

The 1st element in the path
(i.e. AUTHOR) is repeated

Lockstep column traversal

example on one column

AUTHOR

Dumas (0, 1)
Yrsa Sigurdardottir (0, 1)
Tina Flecken (1, 1)
Elma Klein (1, 1)
→ NULL (0, 0)

BOOK 1

AUTHOR: Dumas

BOOK 2

AUTHOR: Yrsa Sigurdardottir

AUTHOR: Tina Flecken

AUTHOR: Elma Klein

BOOK 3

R = 0

We have a new record

Lockstep column traversal

example with two columns

PRICE.EUR

→19 (0, 2)
NULL (0, 0)
12 (0, 2)
11 (1, 2)

PRICE.USD

→20 (0, 2)
NULL (0, 0)
NULL (0, 1)
NULL (1, 1)

BOOK 1

PRICE

EUR: 19
USD: 20

Lockstep column traversal

example with two columns

PRICE . EUR

19 (0, 2)
→ NULL (0, 0)
12 (0, 2)
11 (1, 2)

PRICE . USD

20 (0, 2)
→ NULL (0, 0)
NULL (0, 1)
NULL (1, 1)

BOOK 1

PRICE

EUR: 19
USD: 20

BOOK 2

R = 0

We have a
new record

D = 0

No PRICE

Lockstep column traversal

example with two columns

PRICE.EUR

19 (0, 2)
NULL (0, 0)
→12 (0, 2)
11 (1, 2)

PRICE.USD

20 (0, 2)
NULL (0, 0)
→NULL (0, 1)
NULL (1, 1)

BOOK 1

PRICE

EUR: 19
USD: 20

BOOK 2

BOOK 3

PRICE

EUR: 12

R = 0

We have a
new record

D = 1

No PRICE.USD

Lockstep column traversal

example with two columns

PRICE.EUR

19 (0, 2)
NULL (0, 0)
12 (0, 2)
→11 (1, 2)

PRICE.USD

20 (0, 2)
NULL (0, 0)
NULL (0, 1)
→NULL (1, 1)

BOOK 1

PRICE

EUR: 19
USD: 20

BOOK 2

BOOK 3

PRICE

EUR: 12

PRICE

EUR: 11

R = 1

Repeated PRICE

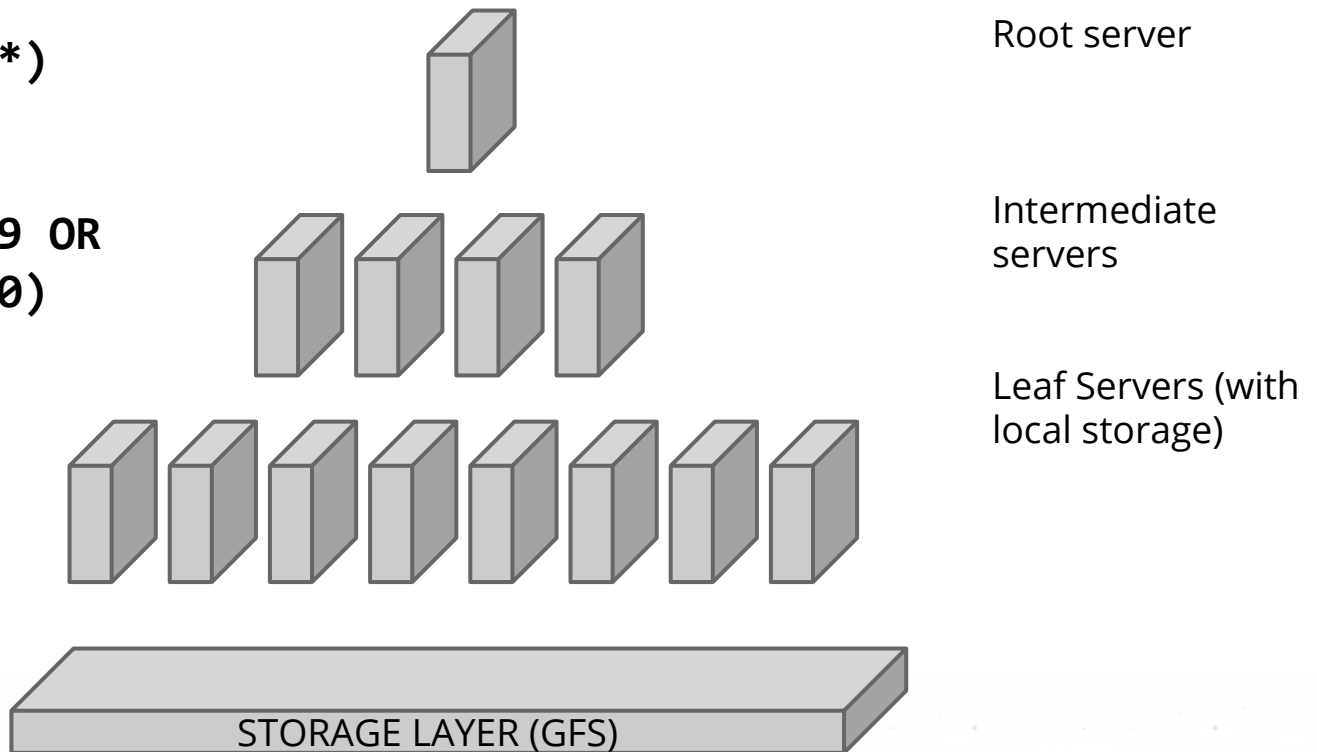
D = 1

No PRICE.USD

Query execution tree

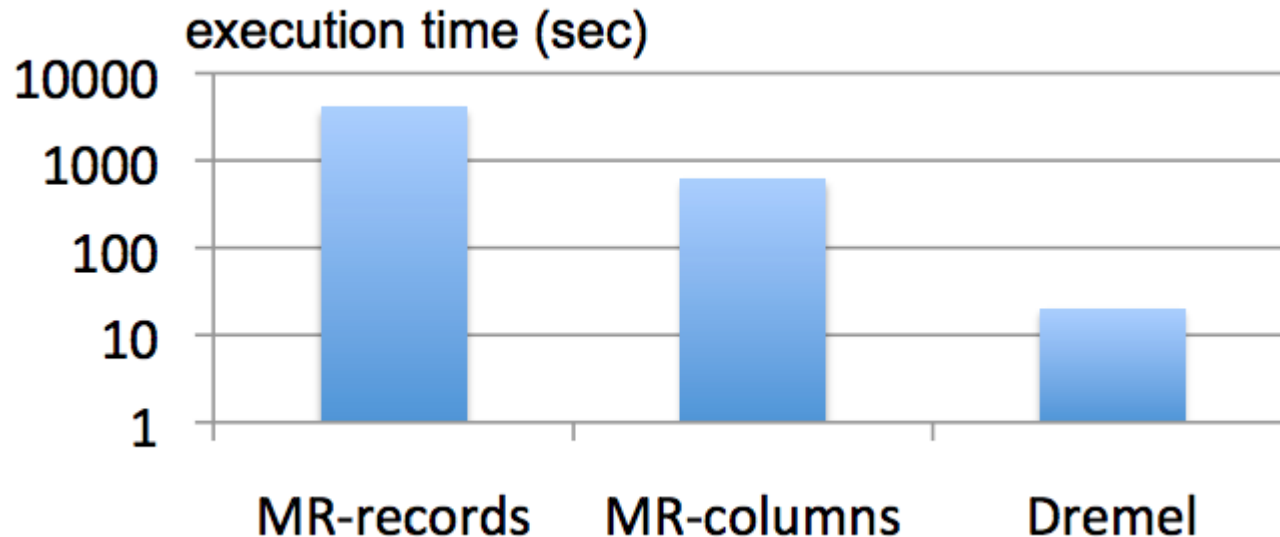
optimised for aggregation queries
using many thousands of servers

```
SELECT Count(*)  
FROM Books  
WHERE  
(PRICE.EUR > 9 OR  
PRICE.USD > 10)
```



Benchmarks

3000 compute nodes, 85 billion records,
87TB (0.5TB read)



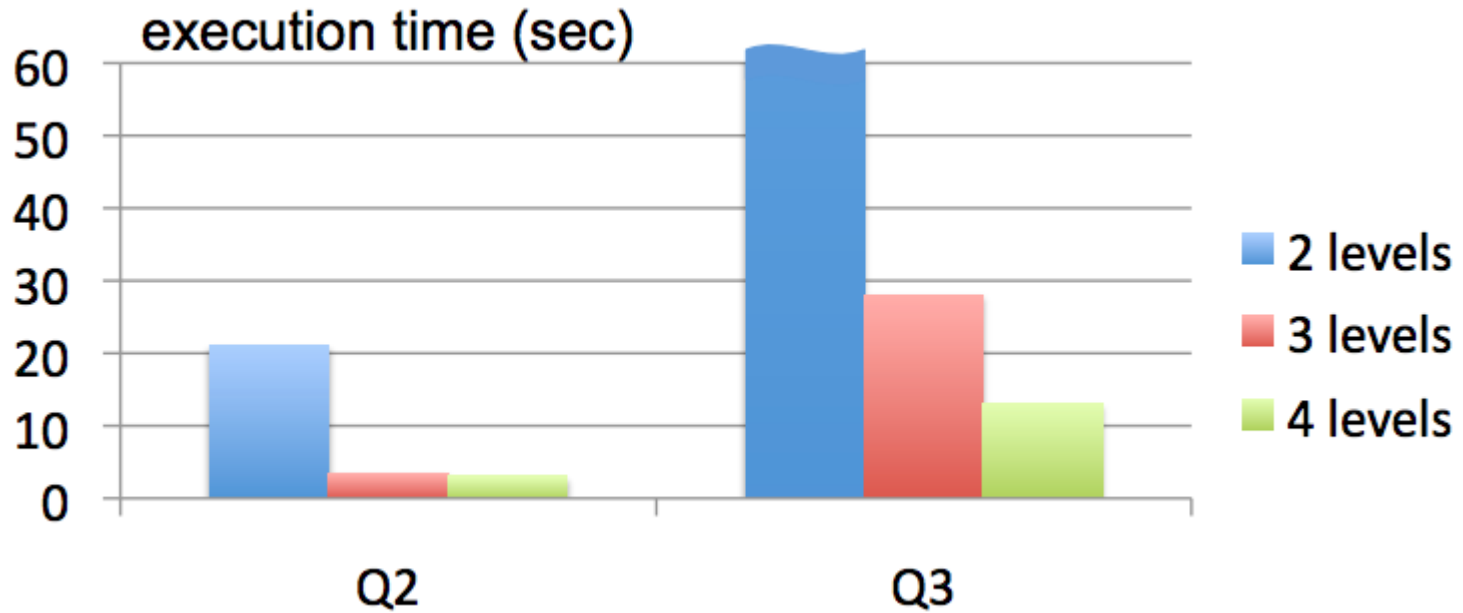
MR = map-reduce

Dremel = BigQuery

The query is `SELECT SUM(CountWords(txtField)) / COUNT(*) FROM T1`

Benchmarks

3000 compute nodes, 24 billion records, 13TB



Q2: SELECT country, SUM(item.amount) FROM T2
GROUP BY country

Q3: SELECT domain, SUM(item.amount) FROM T2 WHERE domain CONTAINS '.net'
GROUP BY domain

Contact



Dirk Primbs

Google Developer Relations

plus.google.com/+DirkPrimbs

