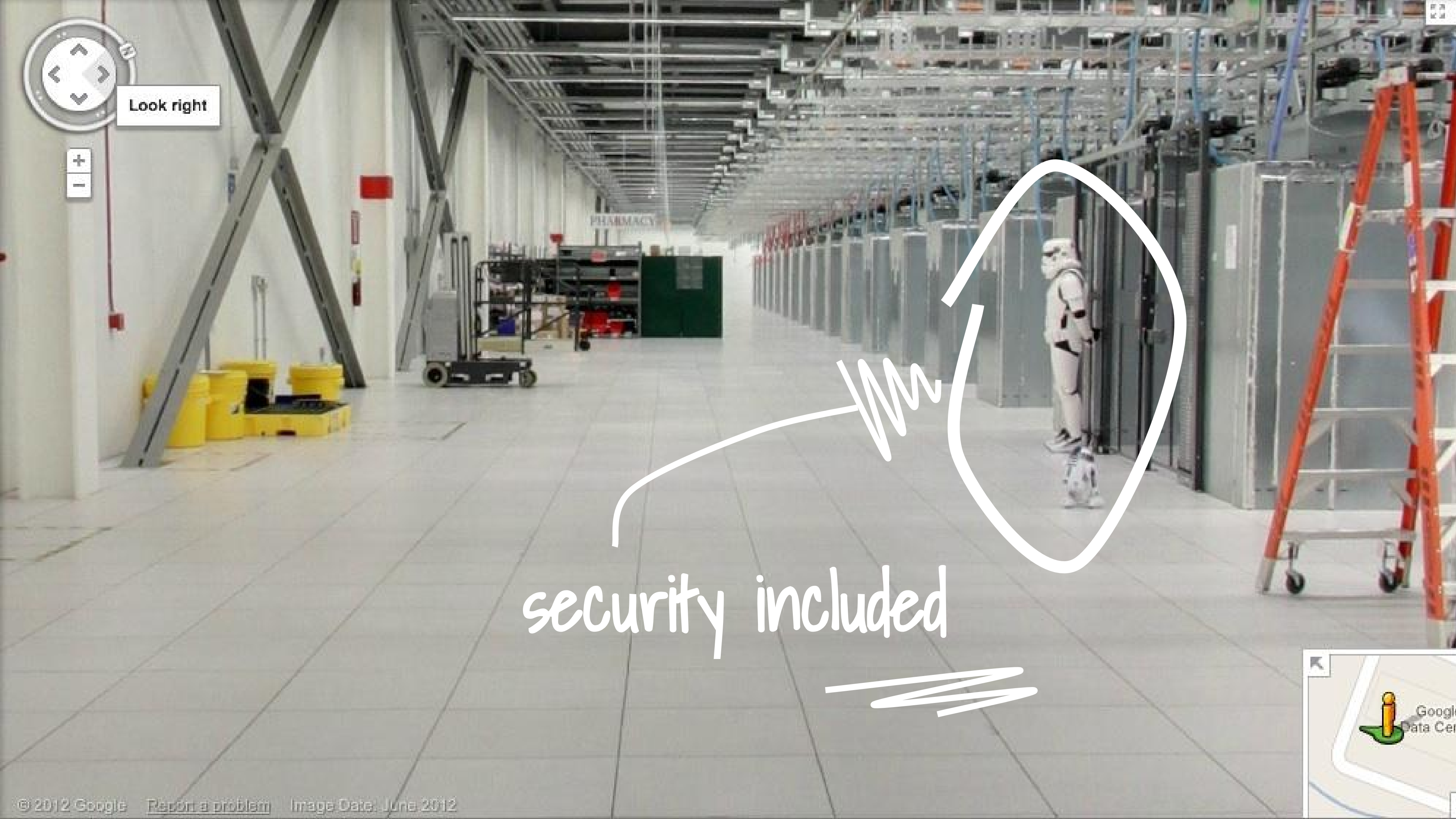
Big Data with the Google Cloud Platform

Nacho Coloma — CTO & Founder at Extrema Sistemas
Google Developer Expert for the Google Cloud Platform
@nachocoloma
http://gplus.to/icoloma

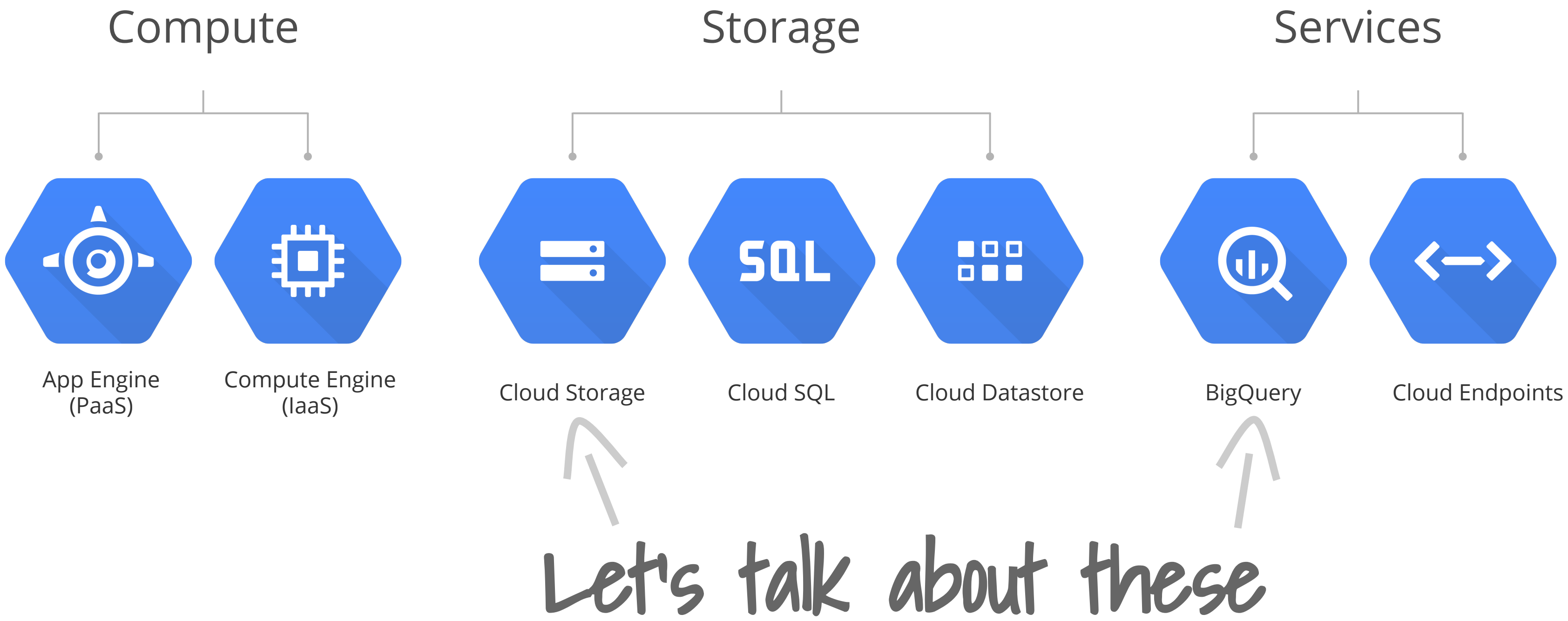
For the past **15 years**, Google has been building the most powerful cloud infrastructure **on the planet.**

Images by Connie Zhou



security included

Google Cloud Platform



Compute

App Engine (PaaS)

Compute Engine (IaaS)

Storage

Cloud Storage

Cloud SQL

Cloud Datastore

Services

BigQuery

Cloud Endpoints

Let's talk about these

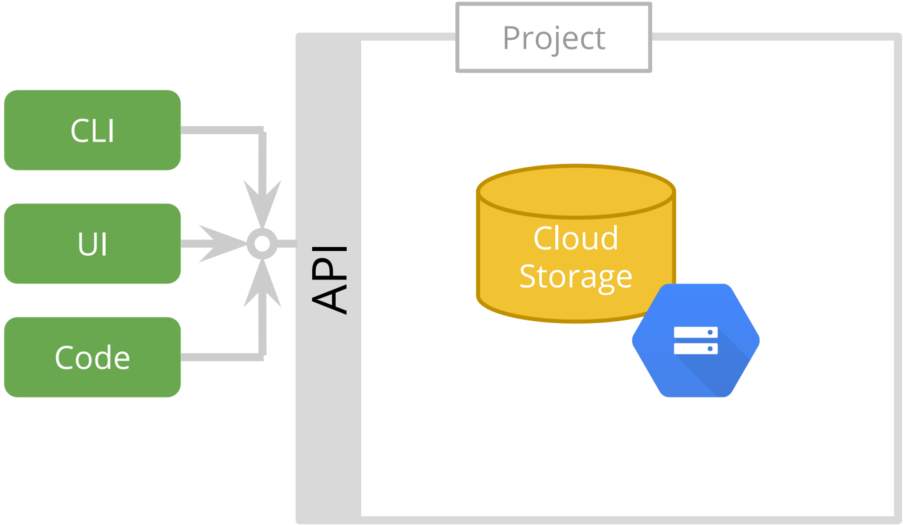
Cloud Storage

```
# create a file and copy it into Cloud Storage
echo "Hello world" > foo.txt
gsutil cp foo.txt gs://<my_bucket>
gsutil ls gs://<my_bucket>

# Open a browser at
https://storage.cloud.google.com/<Your bucket>/<Your Object>
```

Invoking Cloud Storage

CLI: command line
GUI: web console
JSON: REST API



Why go cloud?

Specially if I already have my own data center

Do-it-all yourself

Exploring the Cloud

Do it yourself	IaaS Infrastructure-as-a-Service	PaaS Platform-as-a-Service
Applications	Applications	Applications
Data	Data	Data
Runtime	Runtime	Runtime
Middleware	Middleware	Middleware
O/S	O/S	O/S
Virtualization	Virtualization	Virtualization
Servers	Servers	Servers
Storage	Storage	Storage
Networking	Networking	Networking

🟥 You manage

⬛ Vendor managed

1 minute at Google scale

YouTube 100 hours

1000 new devices

3 million searches

and also…

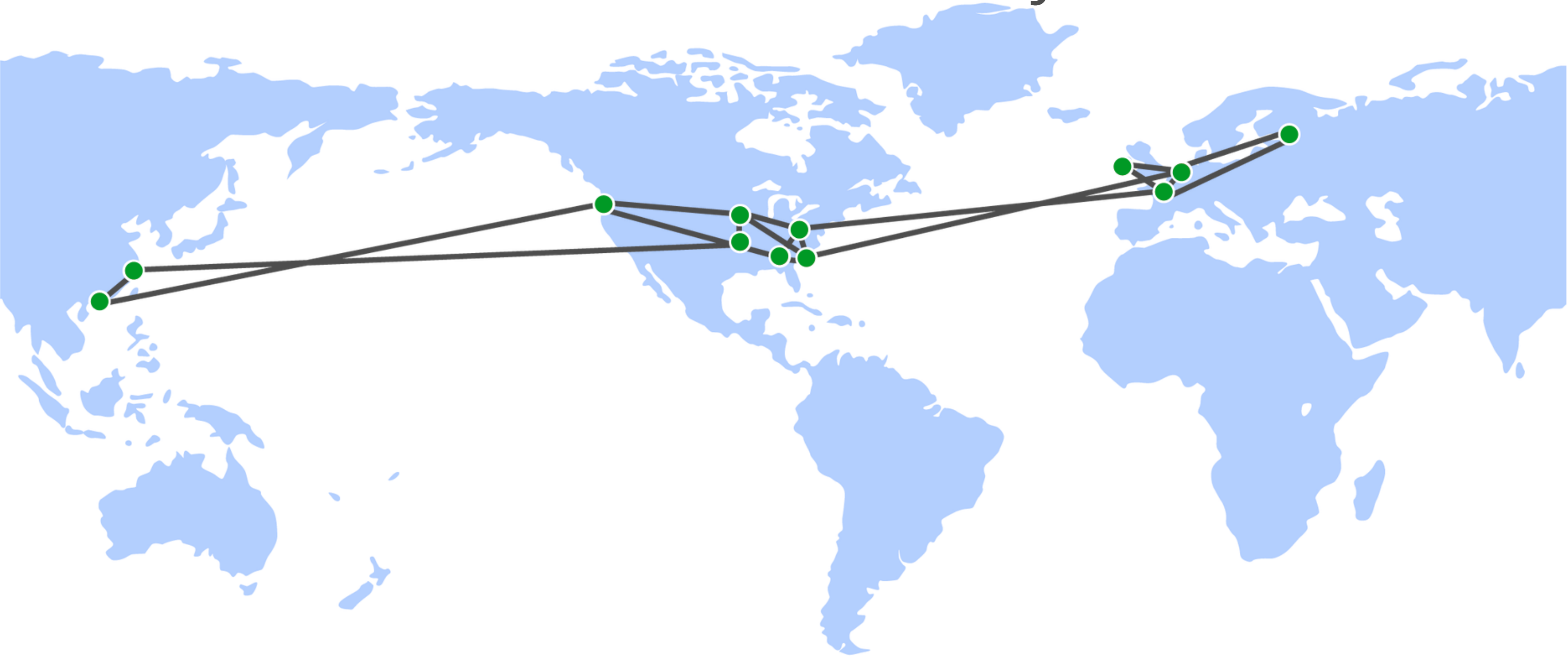
100 million gigabytes

YouTube 1 billion users

1 billion users

1 billion activated devices

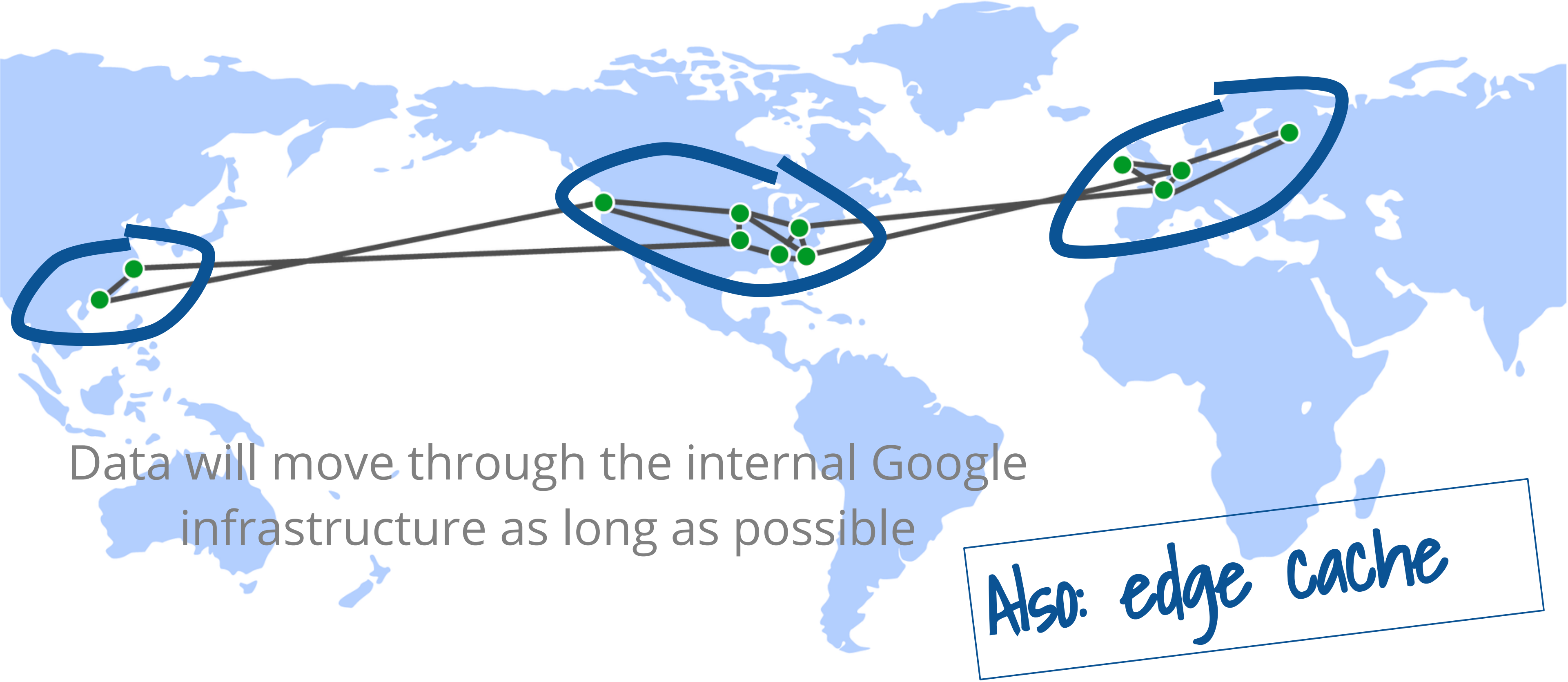
Disaster recovery



Internal bandwidth

Data will move through the internal Google infrastructure as long as possible

Internal bandwidth



Data will move through the internal Google infrastructure as long as possible

Also: edge cache

Cloud Storage: Measure bandwidth

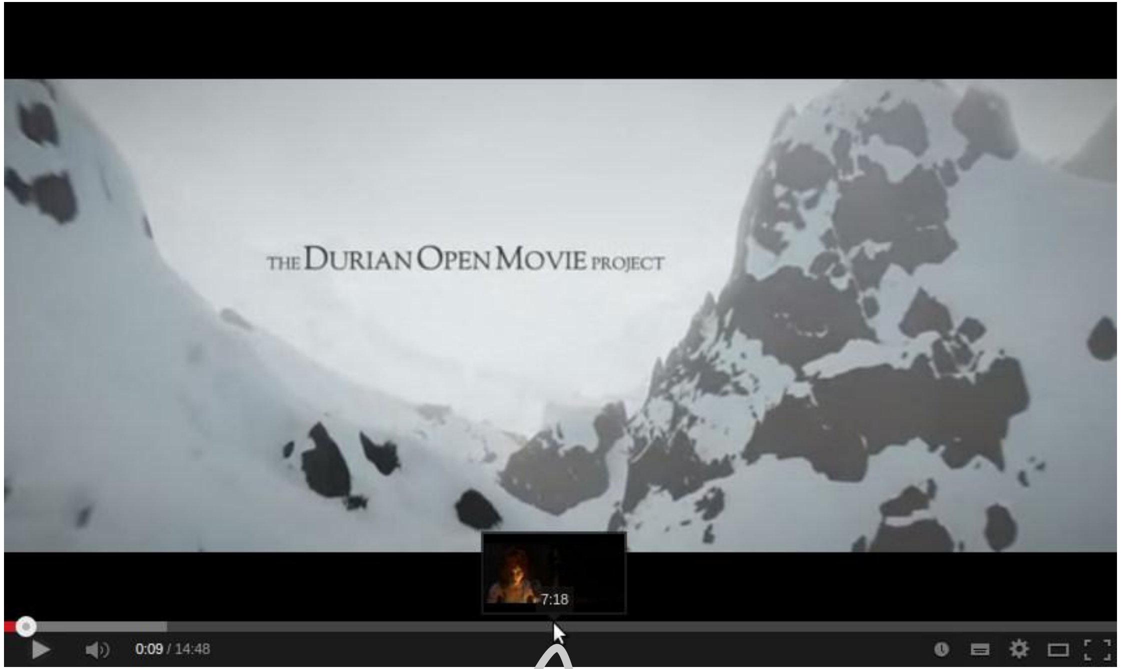
```
# From the EU zone
$ time gsutil cp gs://cloud-platform-solutions-training-exercise-eu/10M-file.txt .
Downloading: 10 MB/10 MB

real   0m10.503s
user  0m0.620s
sys    0m0.456s

# From the US zone
$ time gsutil cp gs://cloud-platform-solutions-training-exercise/10M-file.txt .
Downloading: 10 MB/10 MB

real   0m11.141s
user  0m0.604s
sys    0m0.448s
```

Partial responses



What will happen after clicking here?

Resumable file transfer

Used by gsutil automatically for files > 2MB

Just execute the same command again after a failed upload or download.

Can also be used with the REST API

Parallel uploads and composition

```
# Use the -m option for parallel copying
gsutil -m cp <file1> <file2> <file3> gs://<bucket>

# To upload in parallel, split your file into smaller pieces
$ split -b 1000000 rand-splity.txt rand-s-part-
$ gsutil -m cp rand-s-part-* gs://bucket/dir/
$ rm rand-s-part-*
$ gsutil compose gs://bucket/rand-s-part-* gs://bucket/big-file
$ gsutil -m rm gs://bucket/dir/rand-s-part-*
```

ACLs

- Google Accounts (by ID or e-mail)
- Google Groups (by ID or e-mail)
- Users of a Google Apps domain
- AllAuthenticatedUsers
- AllUsers

Project groups

- Project team members
- Project editors
- Project owners

Durable Reduced Availability (DRA)

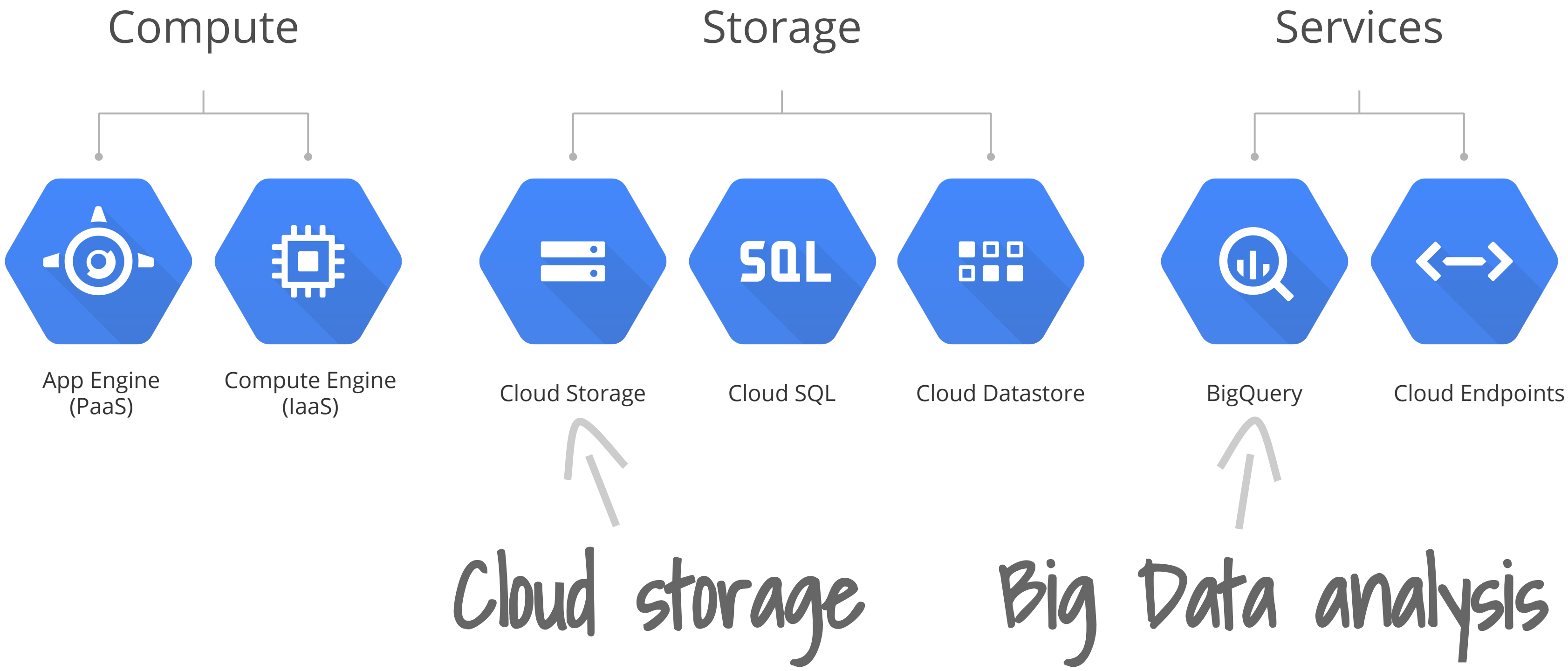
Enables you to store data at lower cost than standard storage (via fewer replicas)

Lower costs
Lower availability
Same durability
Same performance!!!

Object versioning

Buckets can enable object versioning, to undelete files or recover previous versions of your objects.

Google Cloud Platform



Compute

- App Engine (PaaS)
- Compute Engine (IaaS)

Storage

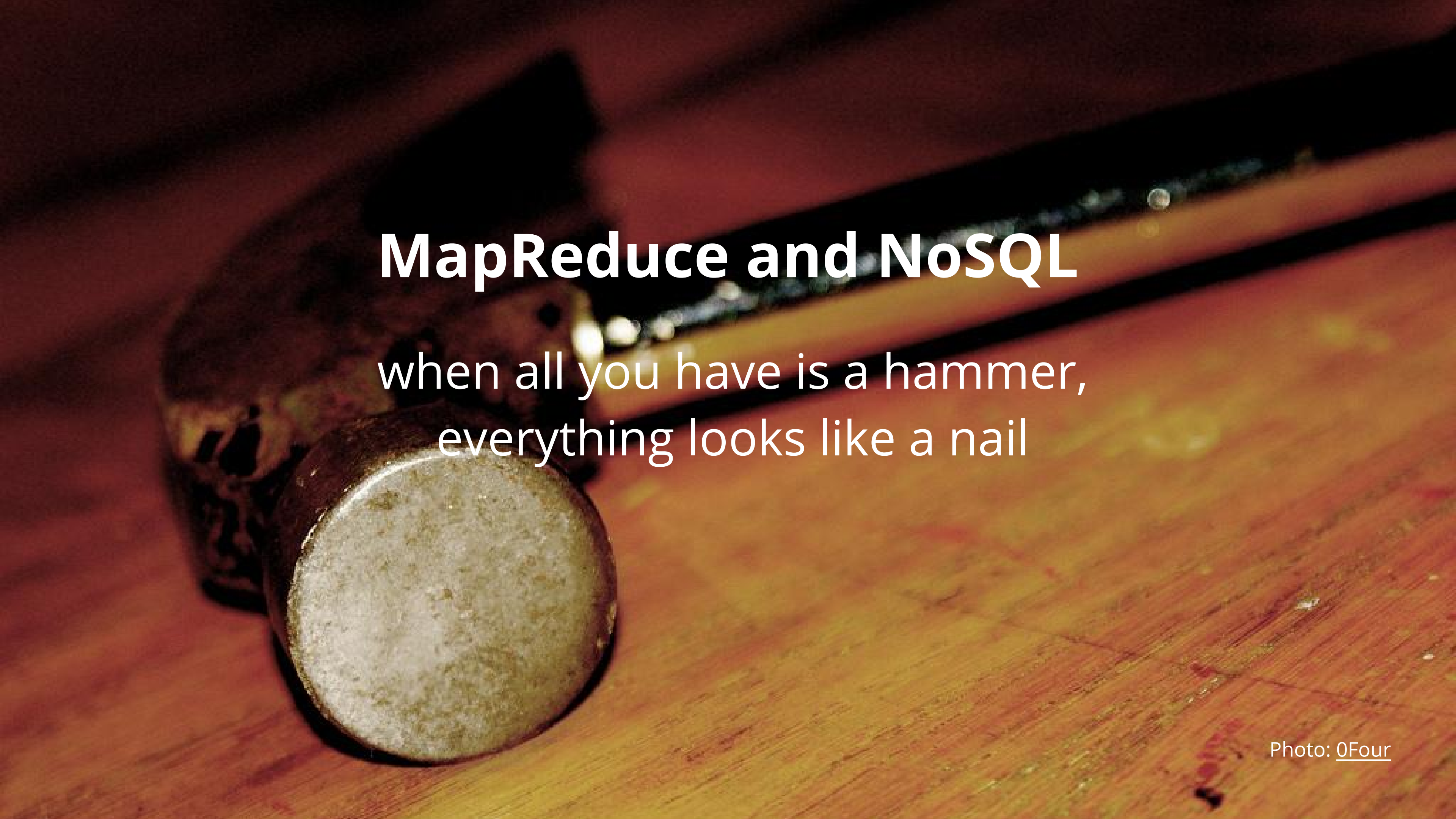
- Cloud Storage
- Cloud SQL
- Cloud Datastore

Services

- BigQuery
- Cloud Endpoints

Cloud storage

Big Data analysis



MapReduce and NoSQL

when all you have is a hammer,
everything looks like a nail

Photo: 0Four

Who is already using AngularJS?

The question that many JavaScript developers are asking

The HTTP Archive

Introduced in 1996

Registers the **Alexa Top 1,000,000 Sites**

About **400GB** of raw CSV data

That's **answers to a lot of questions**

Websites using AngularJS in 2014

	sites using jQuery	sites using AngularJS
Jan	399,258	1297
Feb	423,018	1603
Mar	411,149	1691
Apr	406,239	2004

Not exactly up-to-date, right?

url	rank
http://www.pixnet.net/	122
http://www.zoosk.com/	1256
http://www.nasa.gov/	1284
http://www.udemy.com/	1783
http://www.itar-tass.com/	3277
http://www.virgin-atlantic.com/	3449
http://www.imgbox.com/	3876
http://www.mensfitness.com/	3995
http://www.shape.com/	4453
http://www.weddingwire.com/	4554
http://www.vanityfair.com/	5228
http://www.openstat.ru/	5513

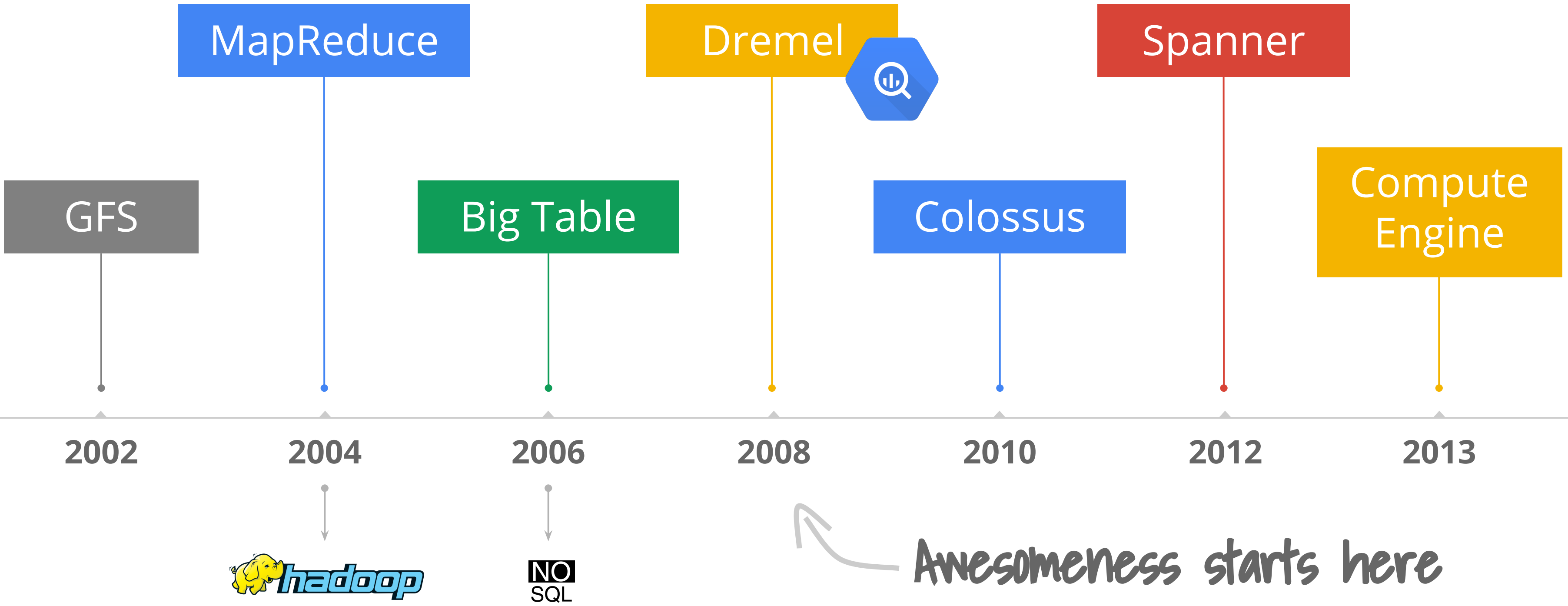
How can we be sure?

```
SELECT pages.pageid, url, pages.rank rank
FROM [httparchive:runs.2014_03_01_pages] as pages
JOIN (
  SELECT pageid
  FROM (TABLE_QUERY([httparchive:runs], 'REGEXP_MATCH(table_id, r"^2014.*requests")'))
  WHERE REGEXP_MATCH(url, r'angular.*\.js')
  GROUP BY pageid
  ) as lib ON lib.pageid = pages.pageid
WHERE rank IS NOT NULL
ORDER BY rank asc;
```

We have a query to validate

Source: http://bigqueri.es

Google innovations in the last twelve years



GFS

2002

MapReduce

2004

hadoop

Big Table

2006

NO SQL

Dremel

2008

Colossus

2010

Spanner

2012

Compute Engine

2013

Awesomeness starts here

Google BigQuery

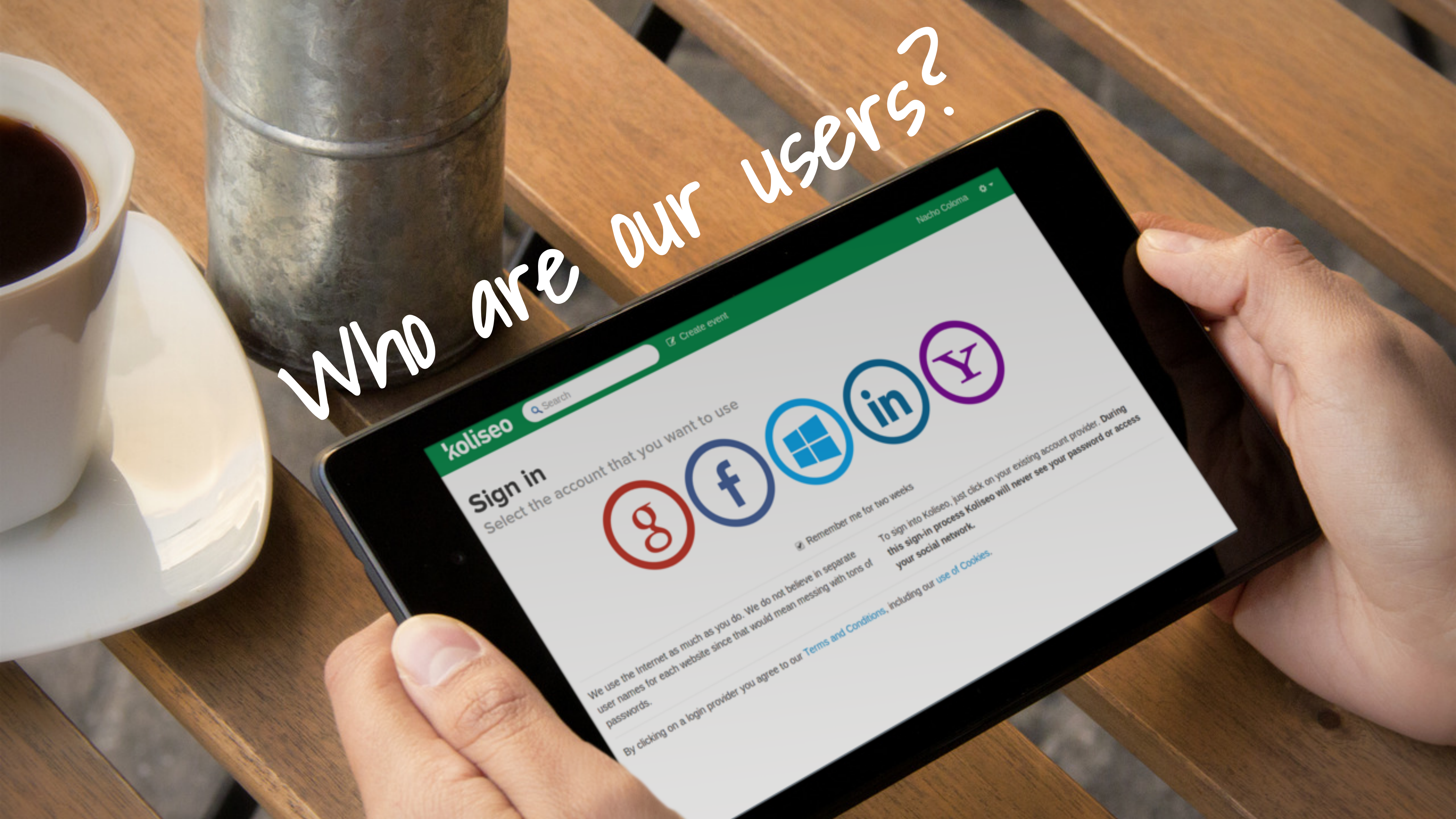
Analyze **terabytes of data in seconds**

Data **imported in bulk** as CSV or JSON

Supports streaming **up to 100K updates/sec per table**

Use the **browser tool**, the **command-line tool** or **REST API**

Who are our users?



Sign in

Select the account that you want to use

☑ Remember me for two weeks

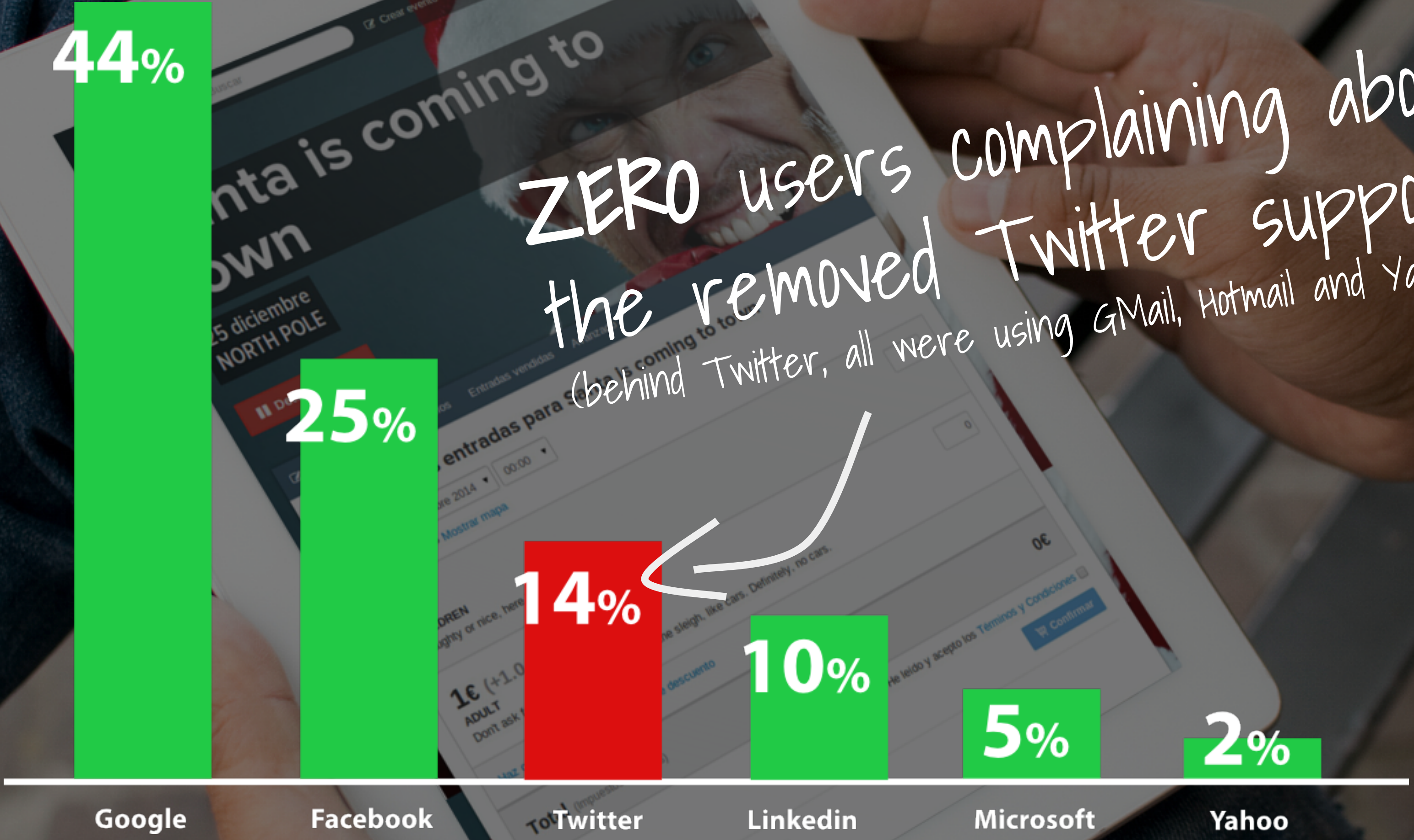
To sign into Koliseo, just click on your existing account provider. **During this sign-in process Koliseo will never see your password or access your social network.**

We use the Internet as much as you do. We do not believe in separate user names for each website since that would mean messing with tons of passwords.

By clicking on a login provider you agree to our Terms and Conditions, including our use of Cookies.

ZERO users complaining about the removed Twitter support

(behind Twitter, all were using GMail, Hotmail and Yahoo!)



BigQuery is a prototyping tool

Answers questions that you need to ask **once in your life**.

Has a flexible interface to **launch queries interactively**, thinking on your feet.

Processes **terabytes of data in seconds**.

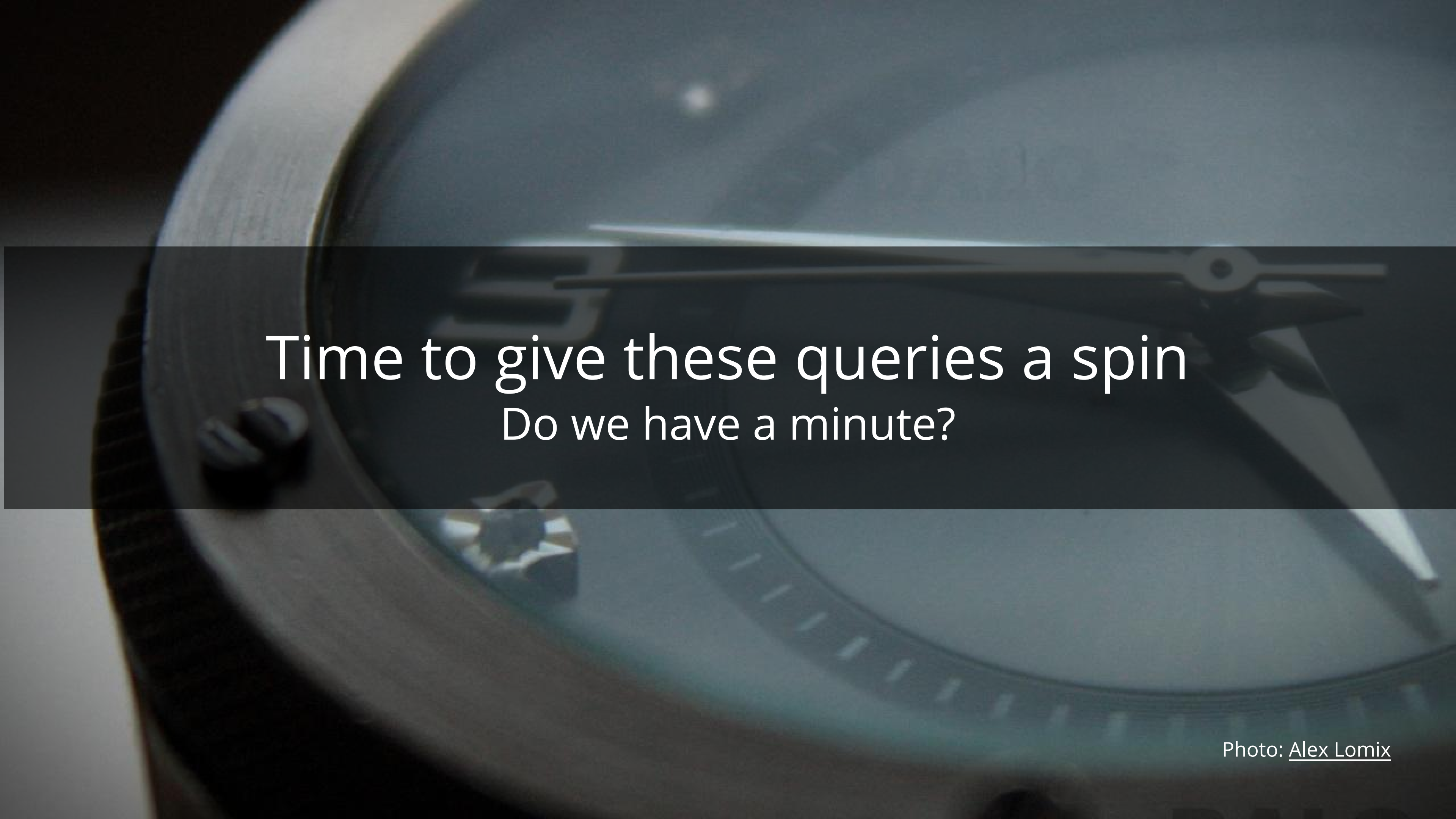
Processes **streaming of data in real time**.

It’s **much easier** than developing Map Reduce manually.

What are the top **100 most active Ruby** repositories on GitHub?

```
SELECT repository_name, count(repository_name) as pushes, repository_description,
repository_url
FROM [githubarchive:github.timeline]
WHERE type="PushEvent"
   AND repository_language="Ruby"
   AND PARSE_UTC_USEC(created_at) >= PARSE_UTC_USEC('2012-04-01 00:00:00')
GROUP BY repository_name, repository_description, repository_url
ORDER BY pushes DESC
LIMIT 100
```

Source: http://bigqueri.es/t/what-are-the-top-100-most-active-ruby-repositories-on-github/9



Time to give these queries a spin

Do we have a minute?

Photo: Alex Lomix

Much more flexible than SQL

Multi-valued attributes

```
    lived_in: [
        { city: ‘La Laguna’, since: ‘19752903’ },
        { city: ‘Madrid’, since: ‘20010101’ },
        { city: ‘Cologne’, since: ‘20130401’ }
    ]
```

Correlation and nth percentile

```
    SELECT CORR(temperature, number_of_people)
```

Data manipulation: dates, urls, regex, IP...

Cost of BigQuery

Loading data	Free
Exporting data	Free
Storage	$0.026 per GB/month
Interactive queries	$0.005 per GB processed
Batch queries	$0.005 per GB processed

Not for dashboards: If you need to launch your query frequently, it’s more cost effective to use MapReduce or SQL

Questions?

Nacho Coloma — CTO & Founder at Extrema Sistemas
Google Developer Expert for the Google Cloud Platform
@nachocoloma
http://gplus.to/icoloma