



Big Data

Turning your data problem into a competitive advantage

Ju-kay Kwek
Product Manager

Navneet Joneja
Product Manager

1 hour in 1 minute

Managing data is hard

There is a better way

You can leverage Google's expertise
to put your data to work for you.





Defining Big Data

Practical problems & opportunities

“ How are hotel reservations for Spain from New York compared with this time last year? ”

“ Do we need to adjust our marketing campaign? Where? ”



“ Which users who signed up last quarter, have also advanced at least 3 levels, and purchased an item worth more than \$5? ”

Claritics - mobile & social user analytics



“ By the time we prepared the data, we'd forgotten what the question was...” ”

Large Google advertising customer



Business & IT trends driving Big Data

Opportunities

Data is a core business asset

Increasingly data is out in the Cloud
(e.g. social, CRM)

New things are possible in the Cloud
(unique algorithms, scale)

Greatly increased speed of sharing and
iteration



Business & IT trends driving Big Data

Opportunities

Data is a core business asset

Increasingly data is out in the Cloud
(e.g. social, CRM)

New things are possible in the Cloud
(unique algorithms, scale)

Greatly increased speed of sharing and
iteration

Challenges

Information is growing faster than ability to
leverage it

Tough for Enterprise to capture all the
data they generate

Scaling traditional BI for Big Data can be
hard

Skills: requires IT, analytics, software
development



What does Big Data look like?

Some common characteristics

Structured, semi-structured, unstructured

Millions if not billions of rows

Too large to process on a single machine

Too large to store on a single machine

High rate of growth

More daily

Diverse industries

Retail point of sales transactions

User activity logs (mobile & social)

Mobile telemetry & smart devices

Industrial & manufacturing

Financial trading

Medical research (e.g. genomics)

Movie rendering & production



A Big Data example

Display ads serving logs (1 week)

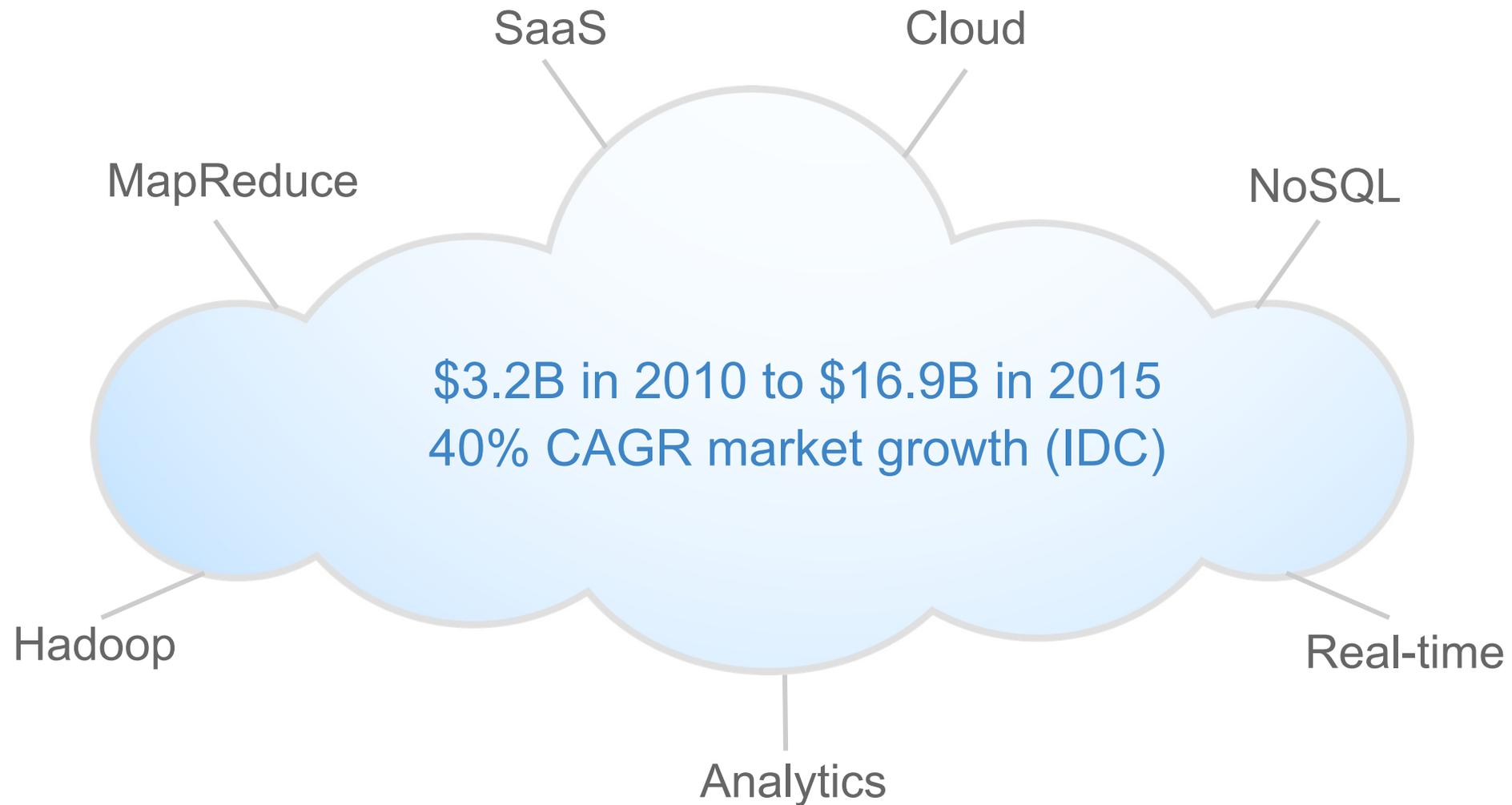
Row	event_type	time	epoch_time	advertiser_id	advertiser	buy_id	buy
1	activity	09-06-2011-13:18:43	1315340323000000	3125202	Google Commerce - NA - Offers - Essence	5753751	Google Offers Q3 Wan
2	activity	09-06-2011-22:49:55	1315374595000000	2878480	Google Chrome: NA - Essence	5748782	Chrome OS Q3 2011 U
3	activity	09-06-2011-15:36:13	1315348573000000	3125202	Google Commerce - NA - Offers - Essence	5753751	Google Offers Q3 Wan
4	activity	09-06-2011-22:35:33	1315373733000000	3226933	Google Chrome: APAC - SEA - Essence	5701556	Chrome Better Web Q3
5	activity	09-06-2011-14:30:23	1315344623000000	2878480	Google Chrome: NA - Essence	5473281	DART Search

300GB
6 billion rows
~900M rows / day

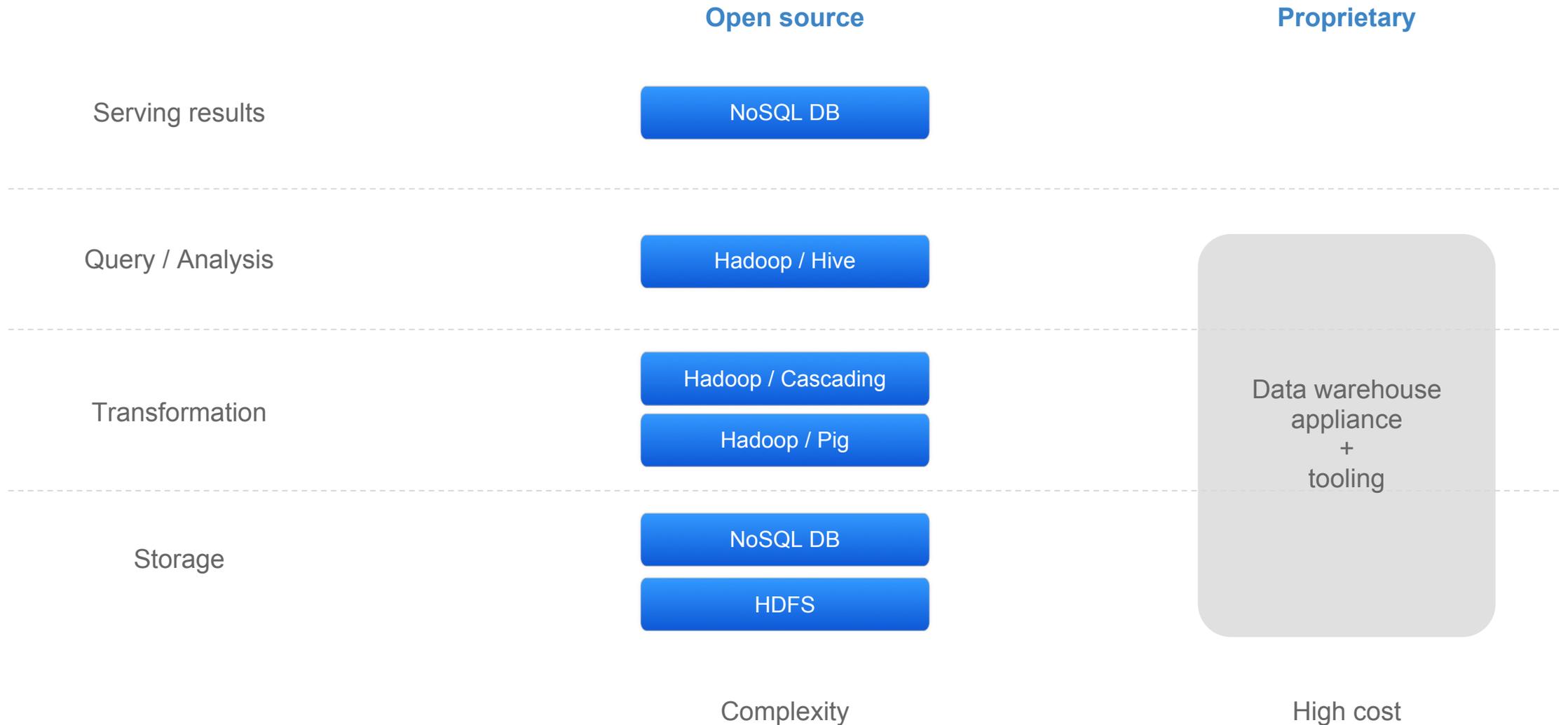
What if you needed to query it?



A look at the Big Data landscape



Emerging Big Data stacks





A better way

Google cloud services for Big Data

Use the cloud



Composable cloud services

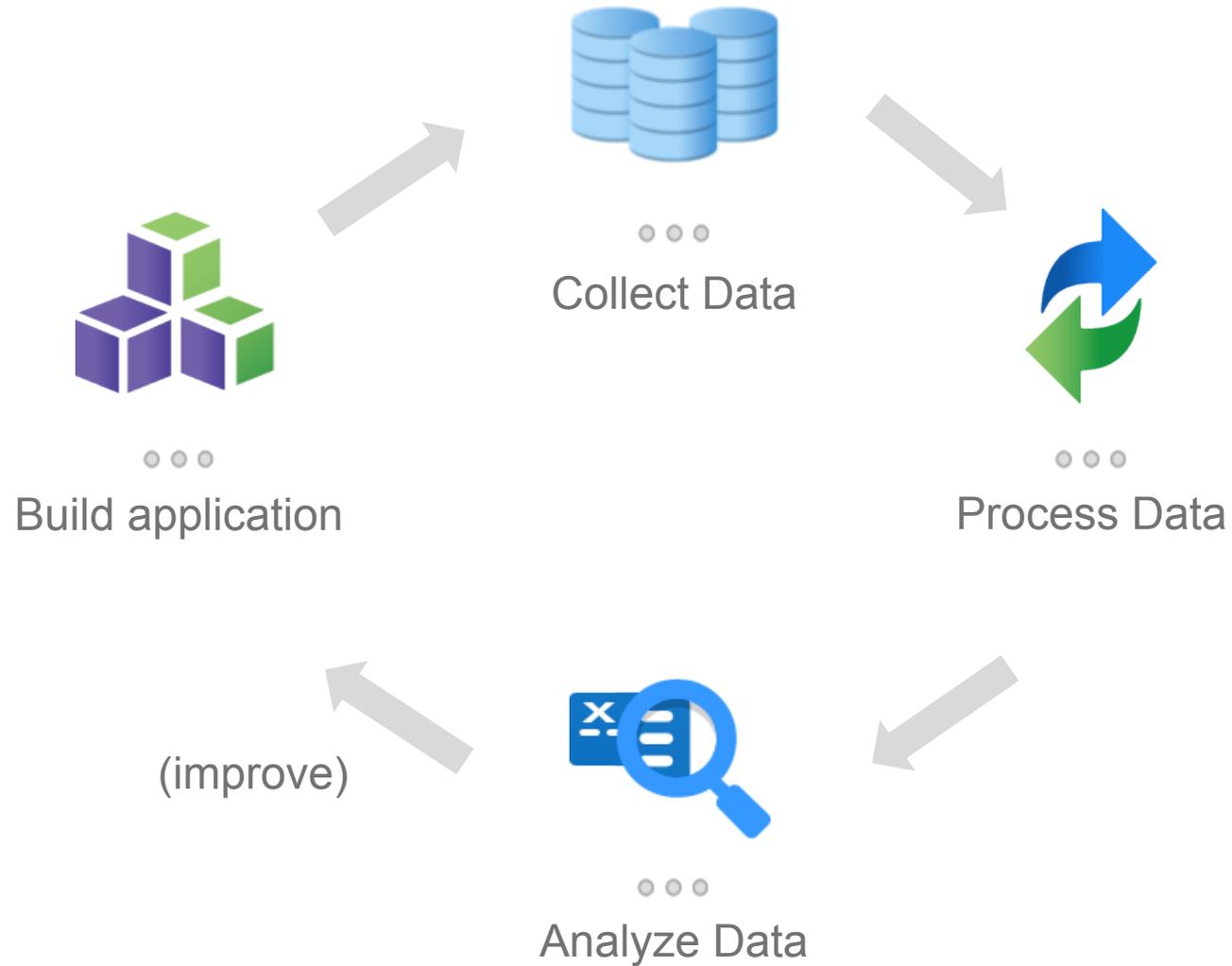
Focus on the solution rather than on the infrastructure

Do new things that weren't possible before

Pay for what you use.



The virtuous cycle of data



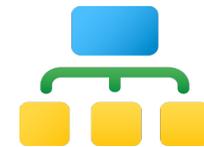
Why Google

Store data with high reliability, redundancy and consistency

Go from Data to Meaning

At Scale

... fast



Google cloud services for data



Google Cloud Storage



Store, access, and manage your data

- Unlimited data
- 5TB per object
- High redundancy
- Simple sharing



Google App Engine



Scalable application development and execution environment

- NoSQL Datastore
- MapReduce
- Long-lived Backends
- Auto-scaling Frontends



Google Compute Engine



Virtual machines

Run arbitrary workloads at scale e.g. Hadoop



Google BigQuery

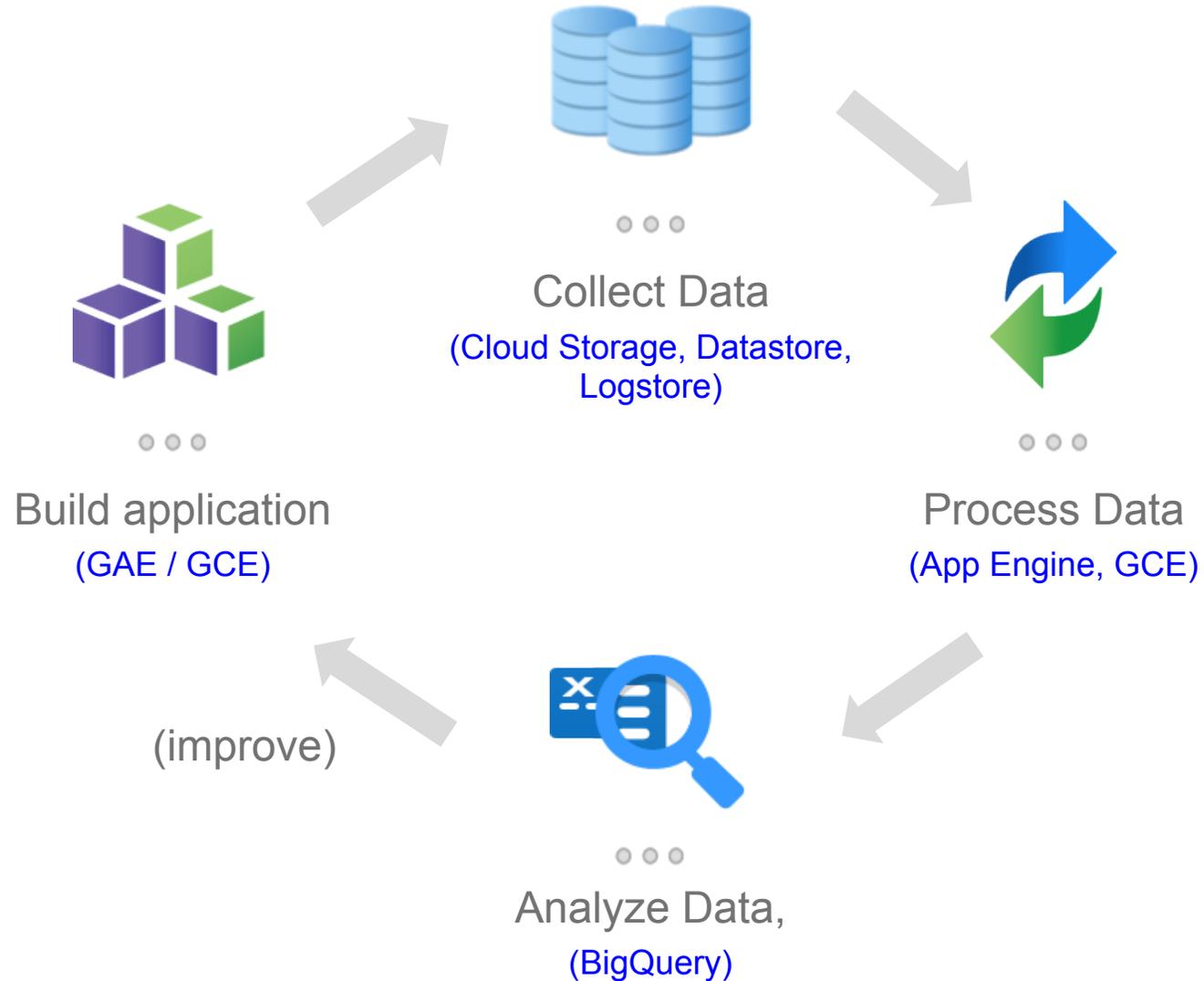


Analyze terabytes of data in seconds

- SQL queries
- Ad hoc response
- Multi-terabyte datasets



The virtuous cycle of data





Build

The Next Generation of Data-Centric Applications

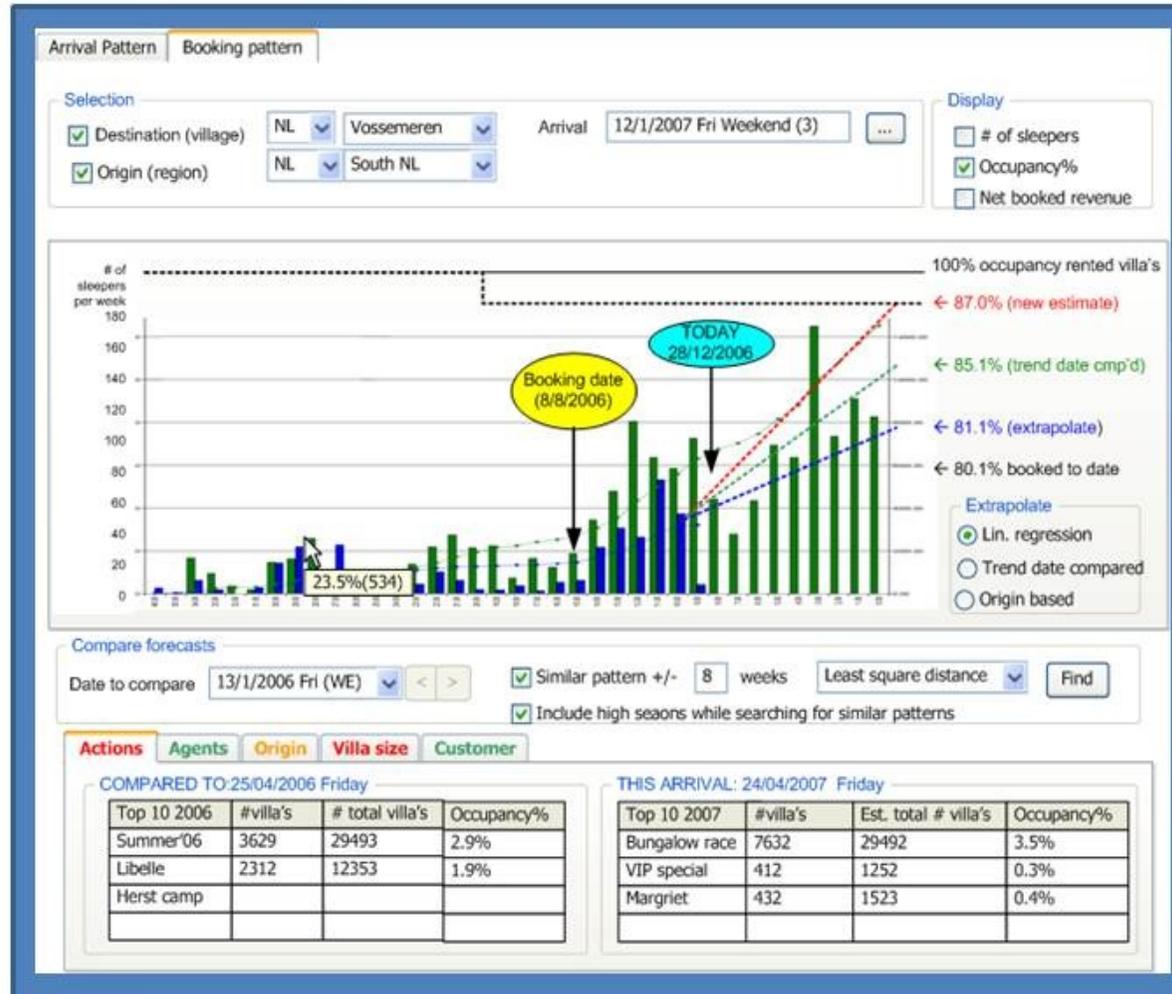
Richard Verhoeff, Co-founder Crystalloids Innovations

About Crystalloids Innovations & CenterParcs

crystalloids
innovations



Solution sketch



← Arrival or Booking
 ← Selection
 ← KPI Measures

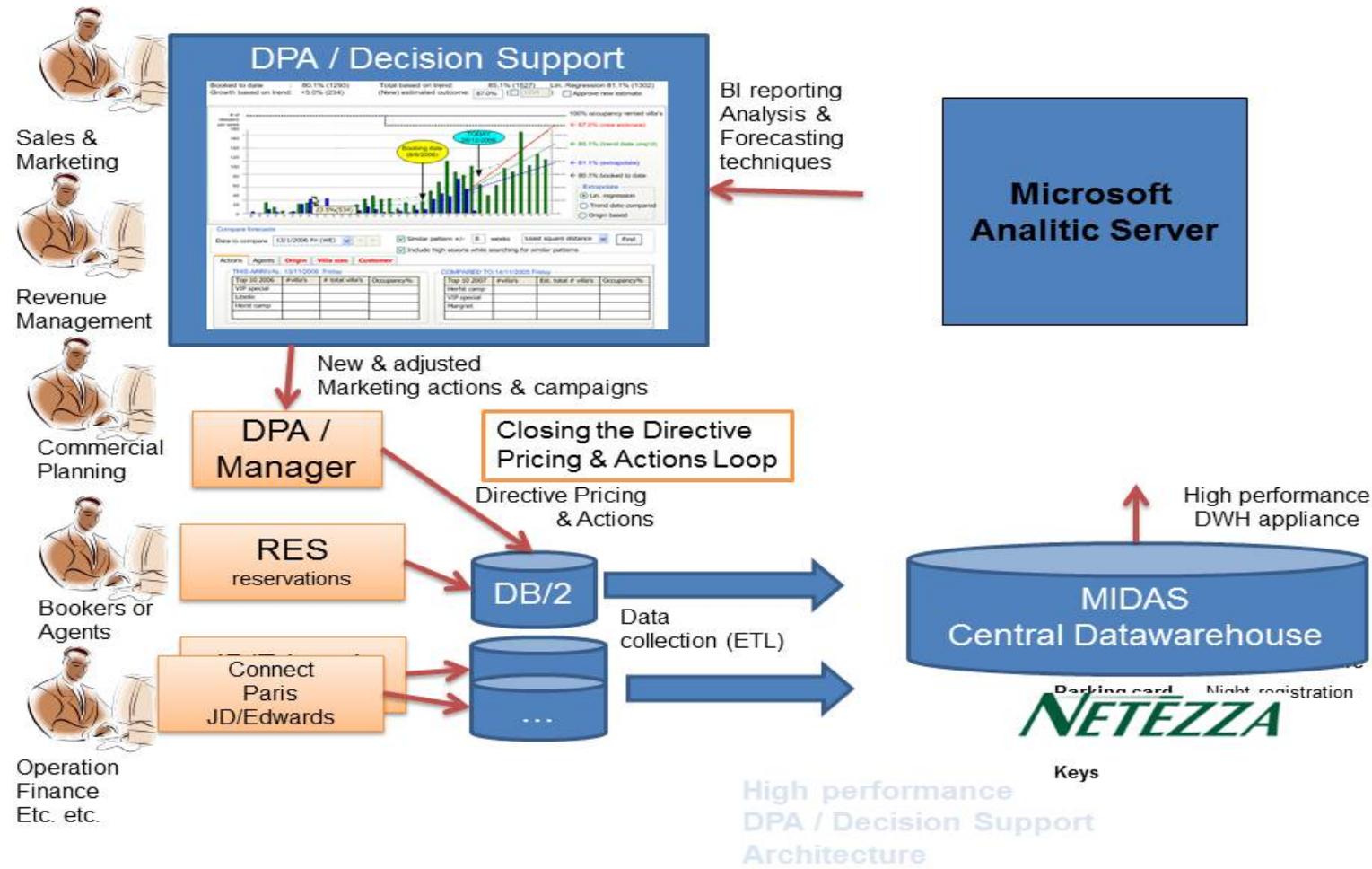
← Booking Pattern
 ← Extrapolations

← Date to compare

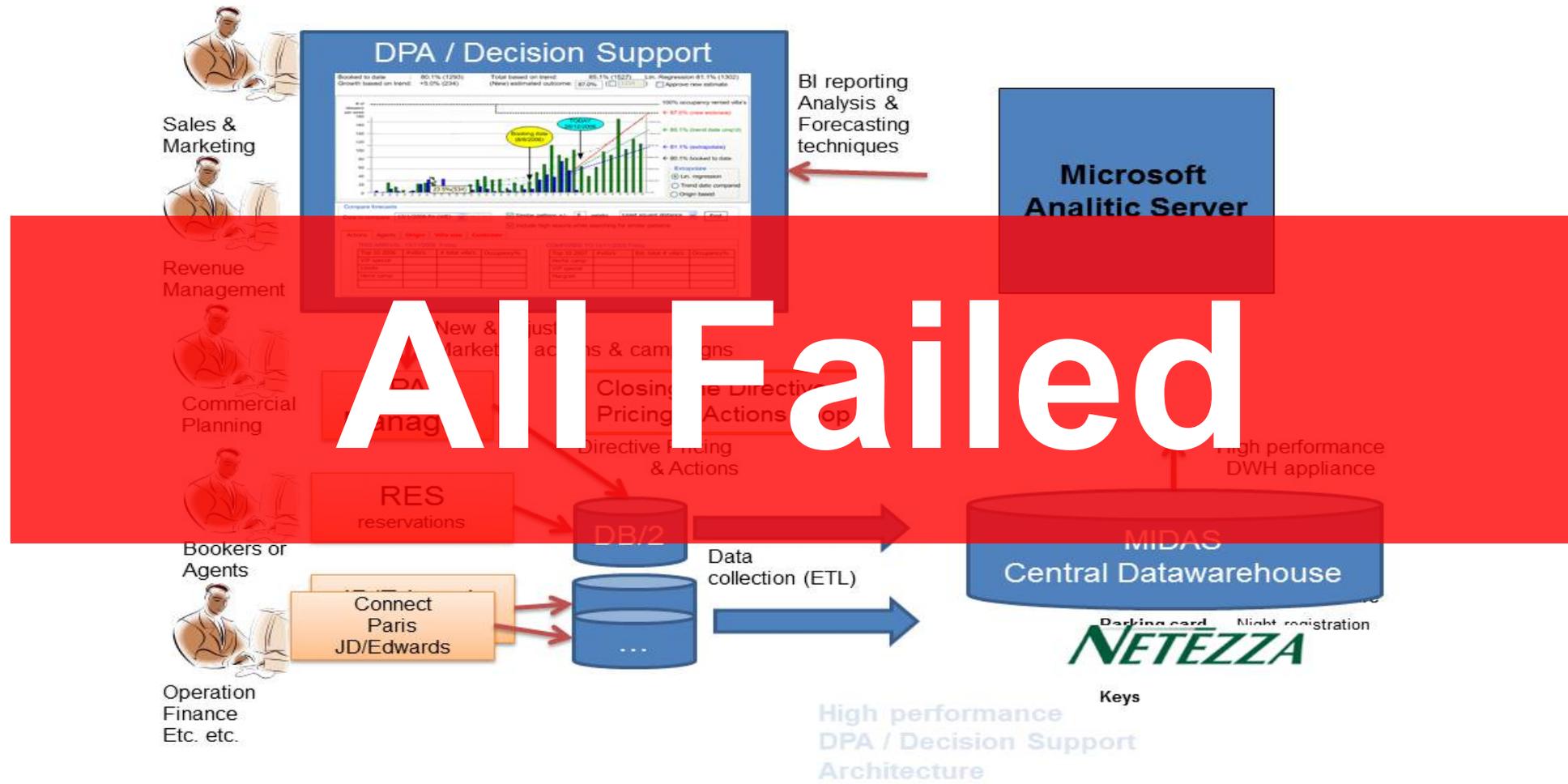
← Demand deviation detection



Complex proofs of concept



Complex proofs of concept



Another Approach...



Focus on business features, via cloud



crystalloids
innovations

New team

- Out-of-the-Box solutions
- Get results quickly

Approach

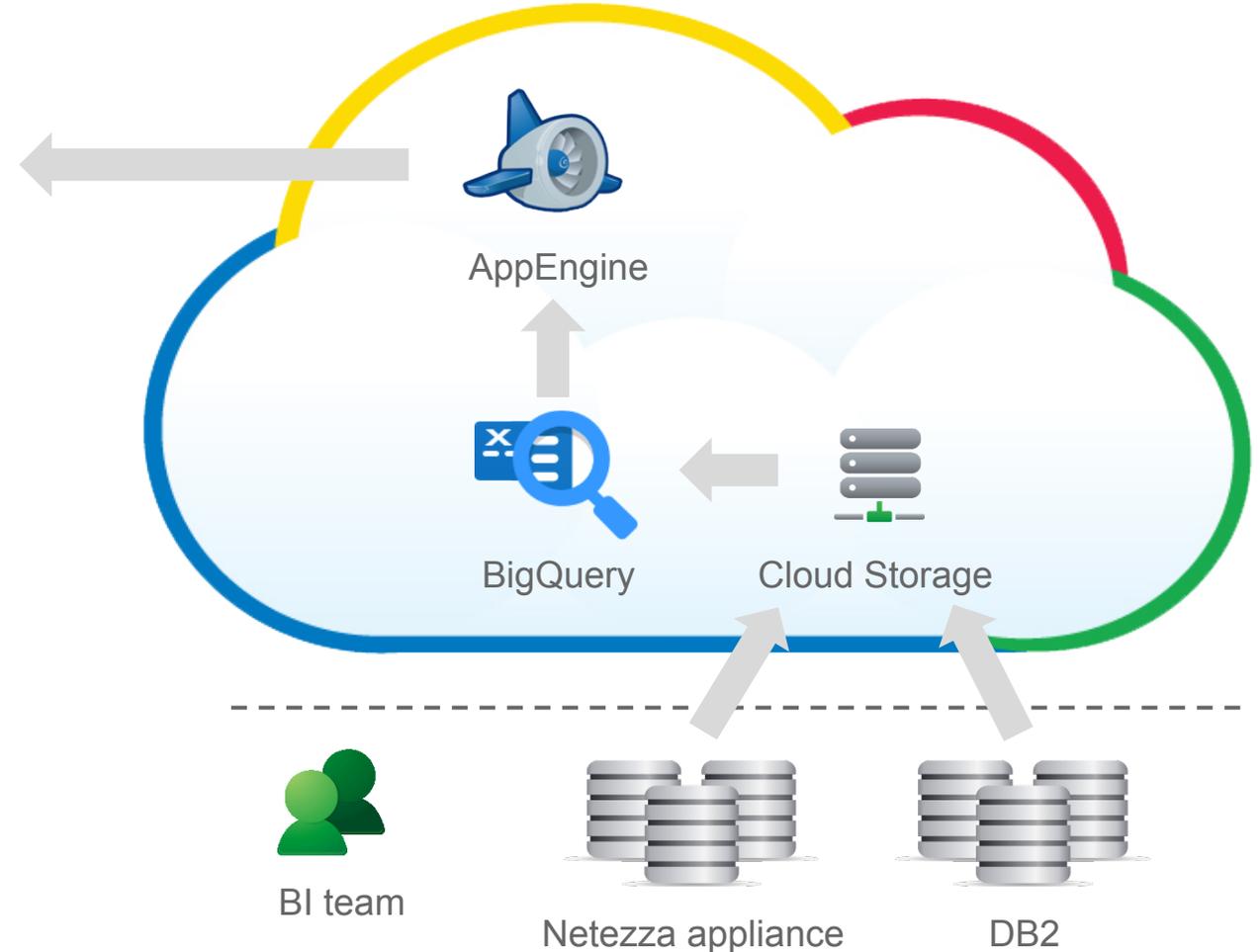
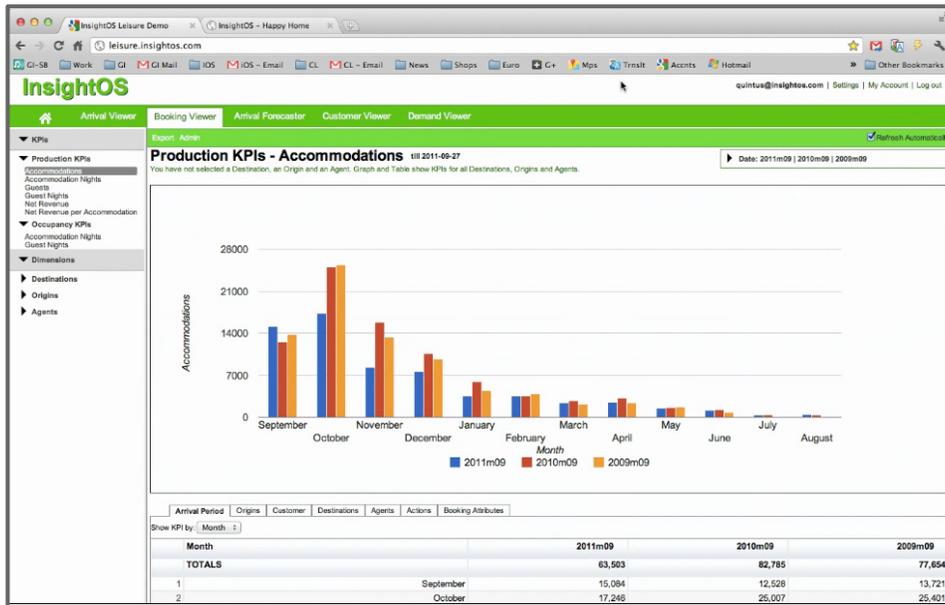
- Focus on Business Features not infrastructure
- Speed, ease of delivery
- Fully-managed cloud solution

Google cloud services

- Google App Engine
- Google Cloud Storage
- Google BigQuery Service
- Google Web Toolkit, Chart Tools, Spreadsheets



Result: cloud-based data mashup




Regional Sales


Analysts


Execs


BI team


Netezza appliance


DB2





Demonstration

demo.insightos.com

Summary



Before

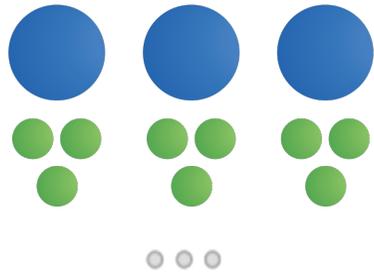
- Minutes or hours to run reports
- Days to present reports to business
- Extremely expensive to scale

After

- Querying millions of rows takes seconds
- Business teams create their own reports using the latest data
- Saved \$800,000 in hardware/software costs



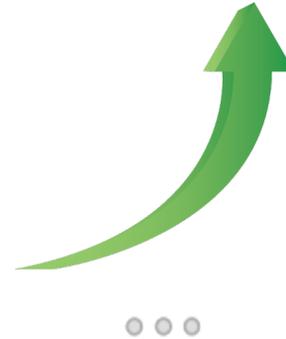
Unstructured Storage: Google Cloud Storage



Multiple Layers of Redundancy



Powered by the Google Network



Unlimited Data; Up to 5 TB per object



Simple Sharing



Google App Engine: serving, data & transformation

Scalable app container

7.5B hits / day
90k QPS

MapReduce

Scale via configuration
80k jobs / day



**Process &
Serve**

Datastore

Largest hosted NoSQL service
Schemaless
Atomic transactions
2 trillion monthly transactions

Logstore

Keep up to a year's worth of logs
Arbitrary logging via API

Store



Google Compute Engine & Hadoop

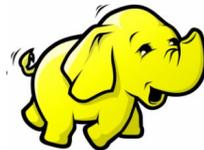
IaaS for large scale data processing
Scalability, performance and value

Goal: deliver #1 platform for custom analytics workloads like hadoop

2 approaches:
Strong partner ecosystem (RightScale, Numerate, Opscode, Puppet Labs, Clickr)
Deploy your own Hadoop cluster



Compute Engine VM



Hadoop



BigQuery

Data cleaning

ETL

Interactive Query Analytics

[302](#): Introducing Google Compute Engine

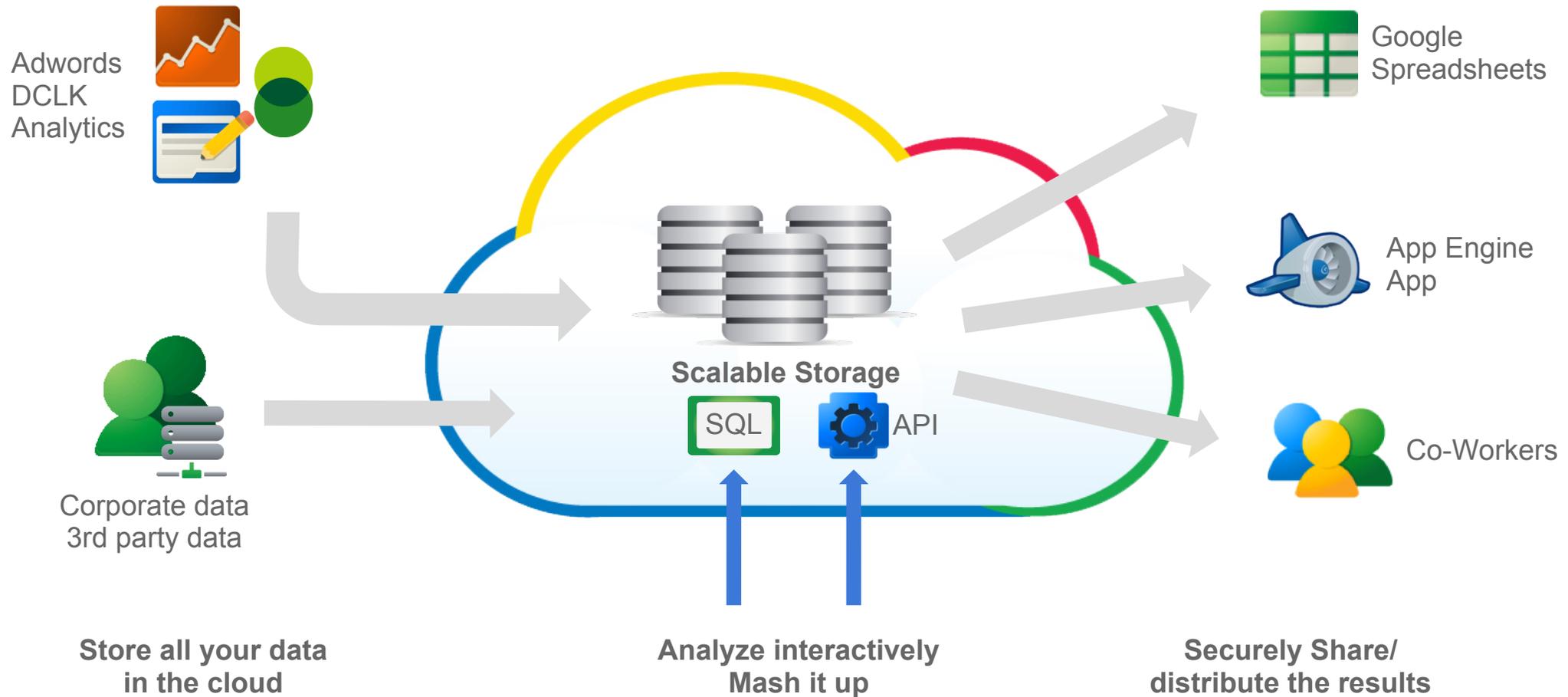
[308](#): Managing Google Compute Engine VMs through App Engine

[313](#): Google Compute Engine -- Technical Details

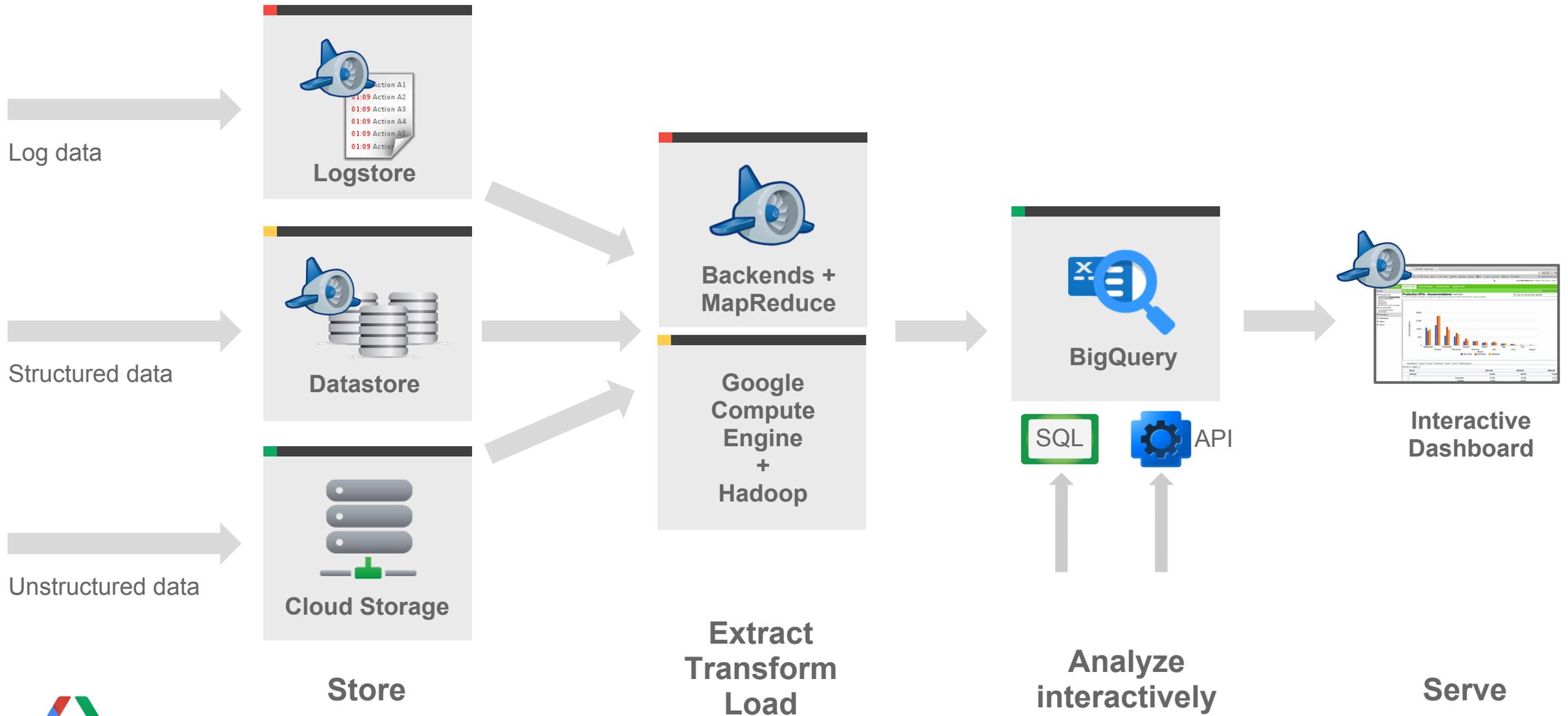


Google BigQuery: fully-managed data analytics in the cloud

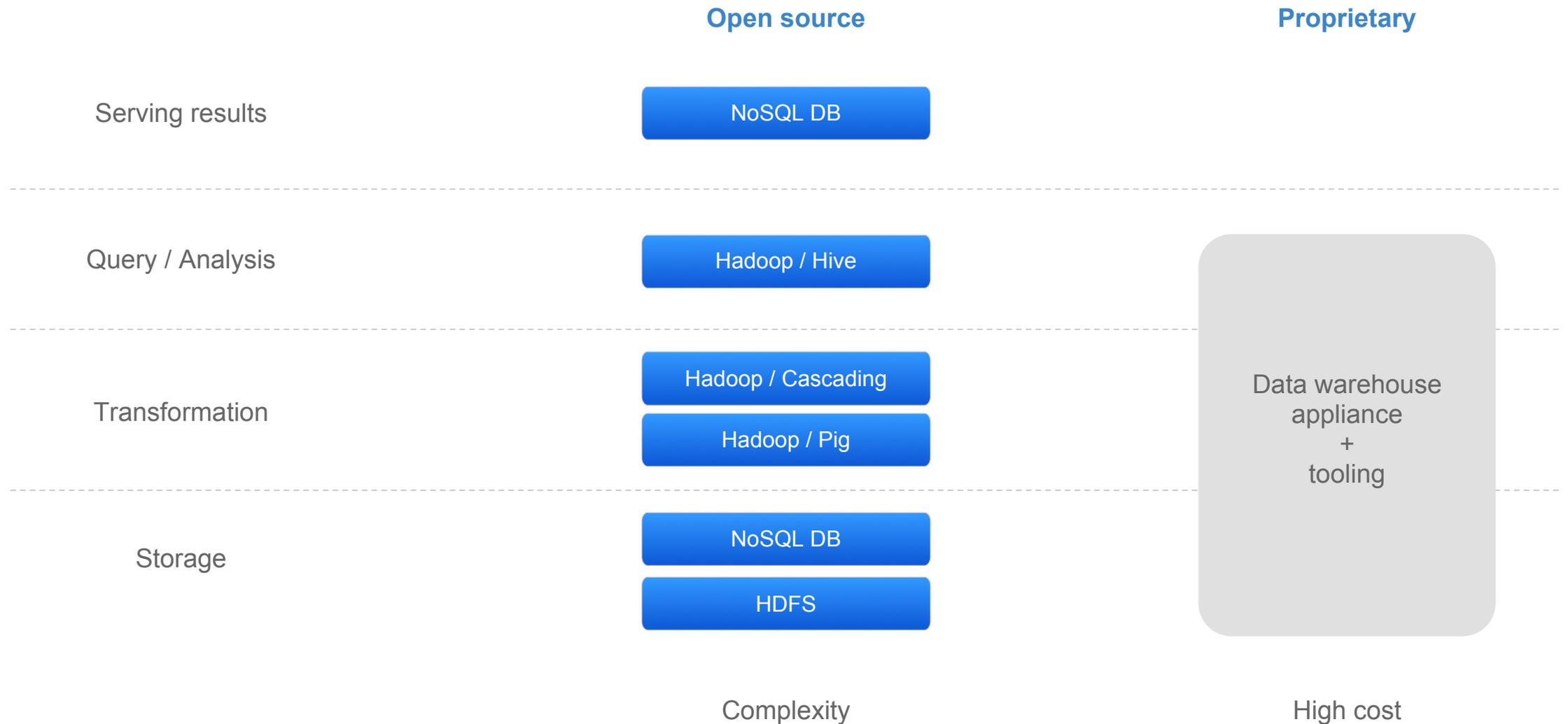
Unlimited storage. Interactive analysis on multi-terabyte datasets



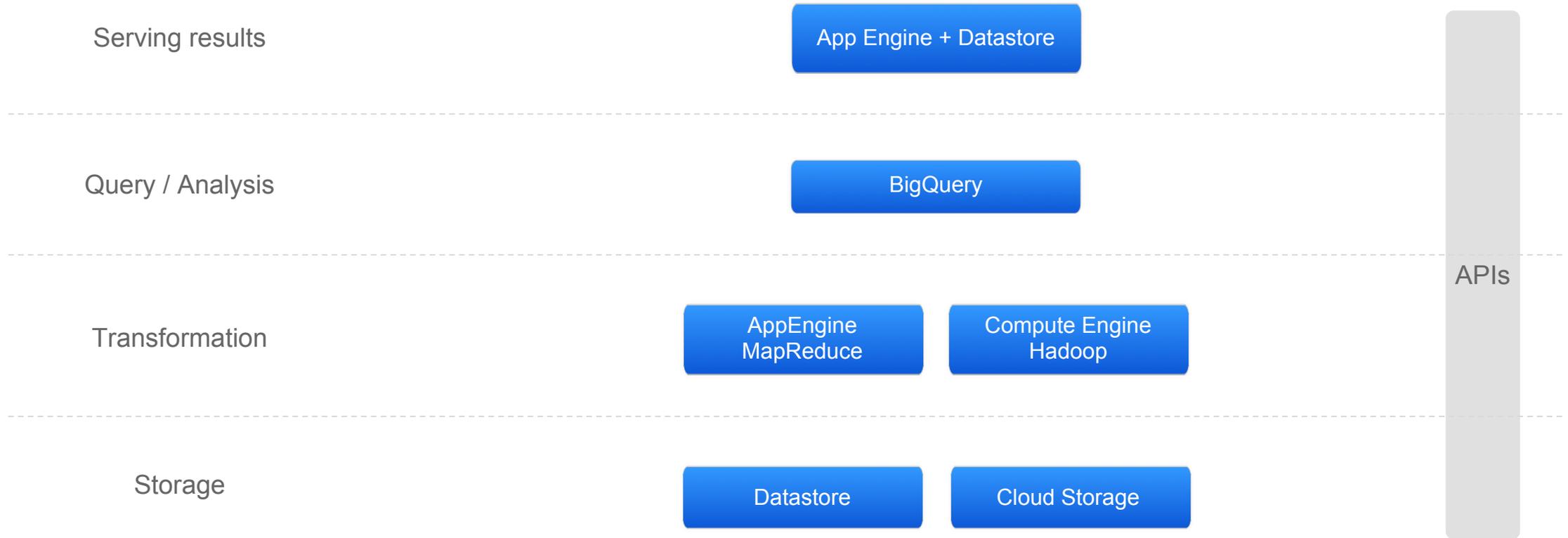
Cloud data analytics pipeline



Emerging Big Data stacks



Google Big Data stack





Building on APIs

Visual Big Data BI using BigQuery

Nicolas Raspal, CTO BIME Analytics

Case study: ad hoc analysis of retail sales

Created in 2009, SaaS Business Intelligence Pioneer

250+ customers worldwide across all industries



Nicolas Raspal
CTO @ We Are Cloud
nicolas@wearecloud.com



Modern Data Analysis and Dashboards

Business Intelligence for the Web Era

Connect (almost) any data source...

Sophisticated but simple pivot table for queries...

Easy to use, beautiful, interactive dashboards...

From the cloud.

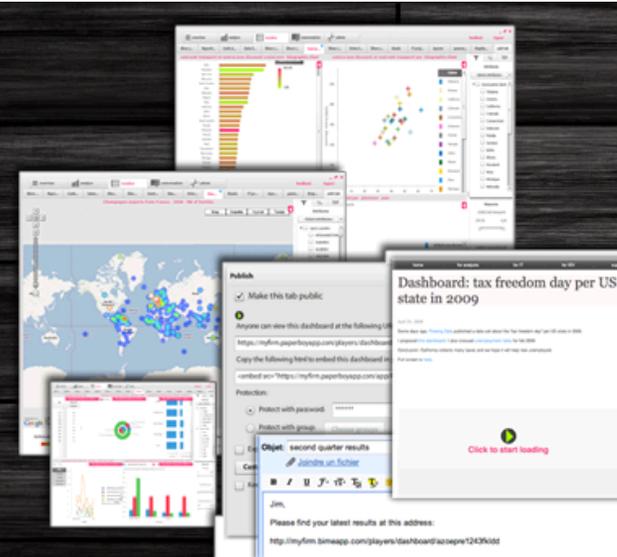


The BigQuery revolution

BI Queries are not normal queries. 95% are aggregate. Have to be fast.

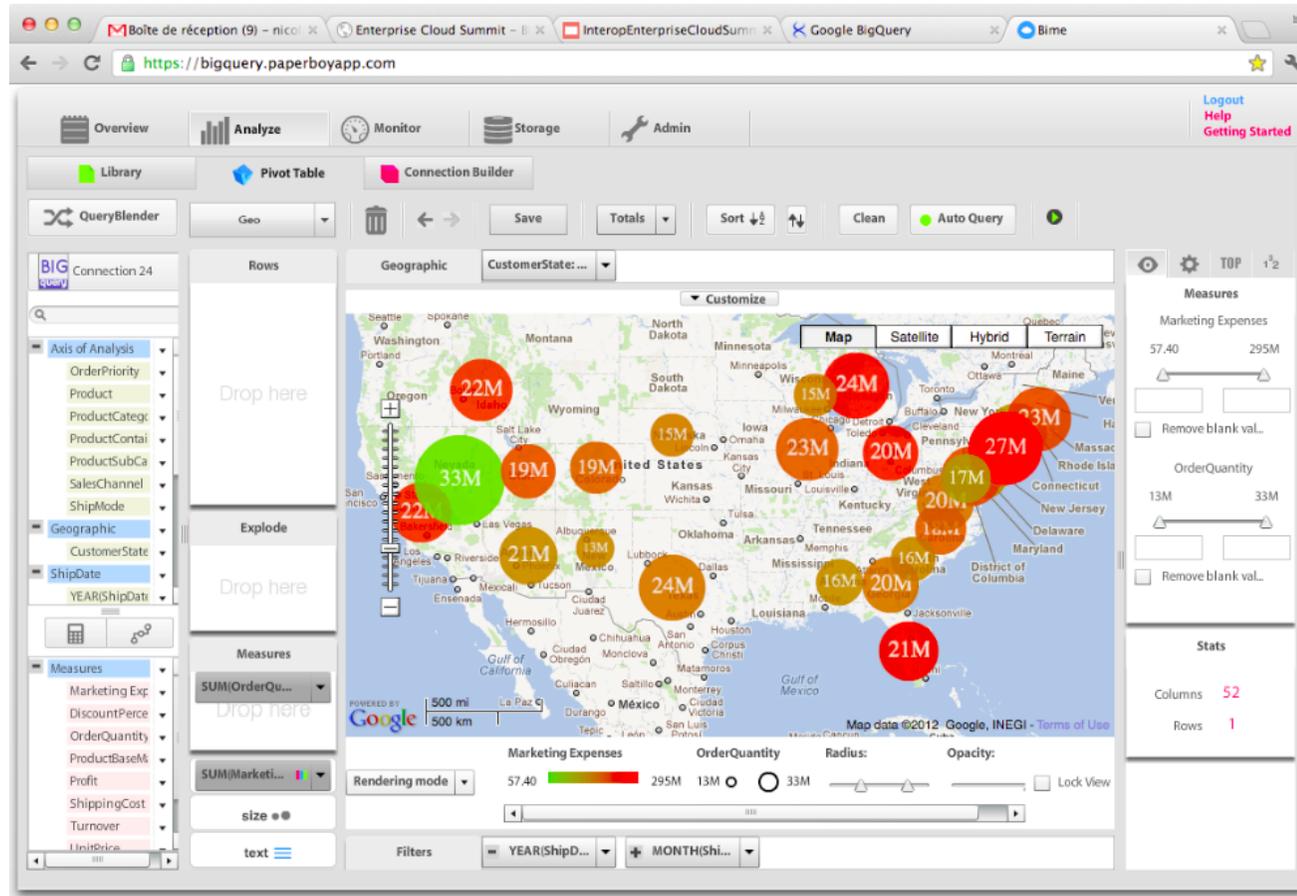
NoSQL, Hadoop, OLAP, relational (column stores or not) DBs are:

- * slow
- * hard to scale
- * complex
- * doesn't support aggregate
- * expensive



Demo: ad hoc analysis of retail sales

Interactively analyze 450M rows of sales data



Why Google cloud services

- Rapidly acquire speed + scalability
- Avoid big on-prem investment
- Simple REST APIs allow rapid iteration



Modern Data Analysis and Dashboards





Data partnerships

Building a Big Data ecosystem

Data technology partners

Storage



ETL



BI / Visualization





Summary

Takeaways for businesses

For the line-of-business

Google provides web-scale services enabling you to start incrementally with Big Data

The more you can ask, the smarter your questions become

Taming *and* freeing your data can turn it into a strategic business asset

For IT & Developers

New school BI: store all the data, analyze & join it later

Cloud enables scalable Big Data pipelines to be set up rapidly

Build apps on data - use the web to set data free



Where to go next...

Related I/O sessions

Building Data Pipelines at Google Scale (Wednesday, June 27 - recorded)

Querying App Engine Logs with BigQuery (Wednesday, June 27 - recorded)

Crunching Big Data with BigQuery (Friday, June 29, 9:00a - 10:00a)

Powering your Application's data using Google Cloud Storage (Friday, June 29, 2:30p)

Developer Sandbox

BIME Analytics

Try it out!

<https://cloud.google.com>



<Thank You!>

For more information, visit <http://cloud.google.com>

jukay@google.com

[#jukaykwek](#)

[+jukay@google.com](tel:+14152531714)

njoneja@google.com

<#>

[+njoneja@google.com](tel:+14152531714)





Google
Developers