Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Big Data
Get ready to be agile & to use multiple solutions

Olivier Flebus

February 16th, 2012

Architecture Week

Olivier Flebus (olivier.flebus@capgemini.com)

- **Managing Enterprise Architect**
- **Agile Community Leader for Capgemini Aerospace & Defence (Toulouse, France)**
- **Twitter @olivierflebus**

bit.ly/flebus

BigData is *really* big

Big data:
the next chapter of
the information revolution

bit.ly/pFV1aw

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

bit.ly/q8CZ6J

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

BIM | the way we see it

Big Data represents a big opportunity and a big reality

bit.ly/o8xB2E

Capgemini's BIM Global Service Line

- Business Information Management is business critical

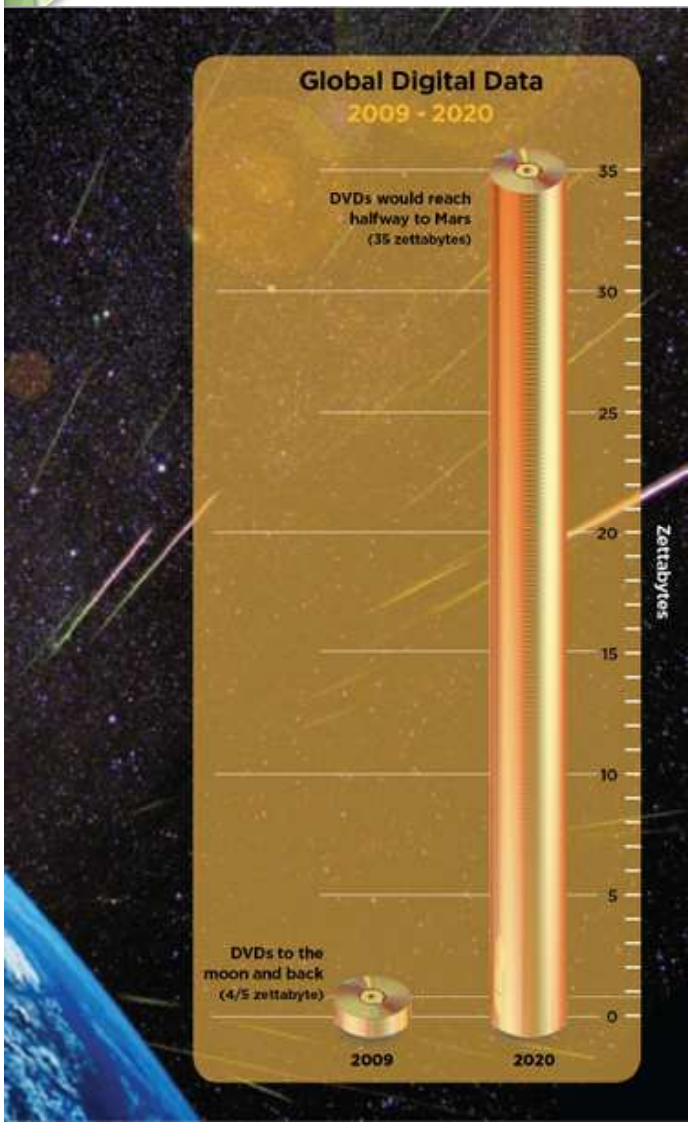
- Turn Data into Competitive advantage through the right Decision-Making

"Drowning in data, Thirsting for knowledge"

bit.ly/nsQgLk

What is BigData ?

1. Volume



"Every two days now we create as much information as we did from the dawn of civilization up until 2003."

Eric Schmidt

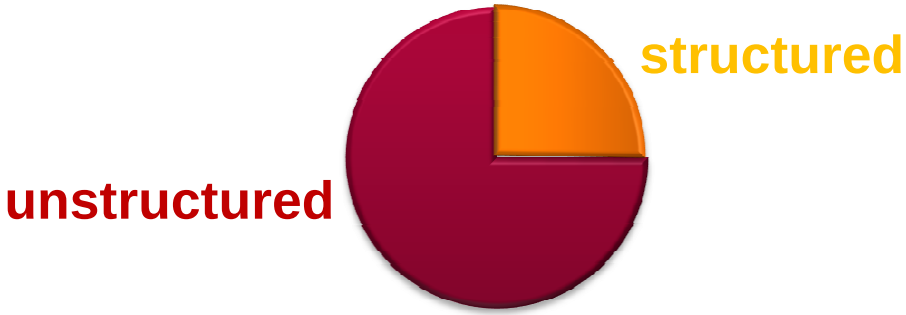
bit.ly/oNyJ2o

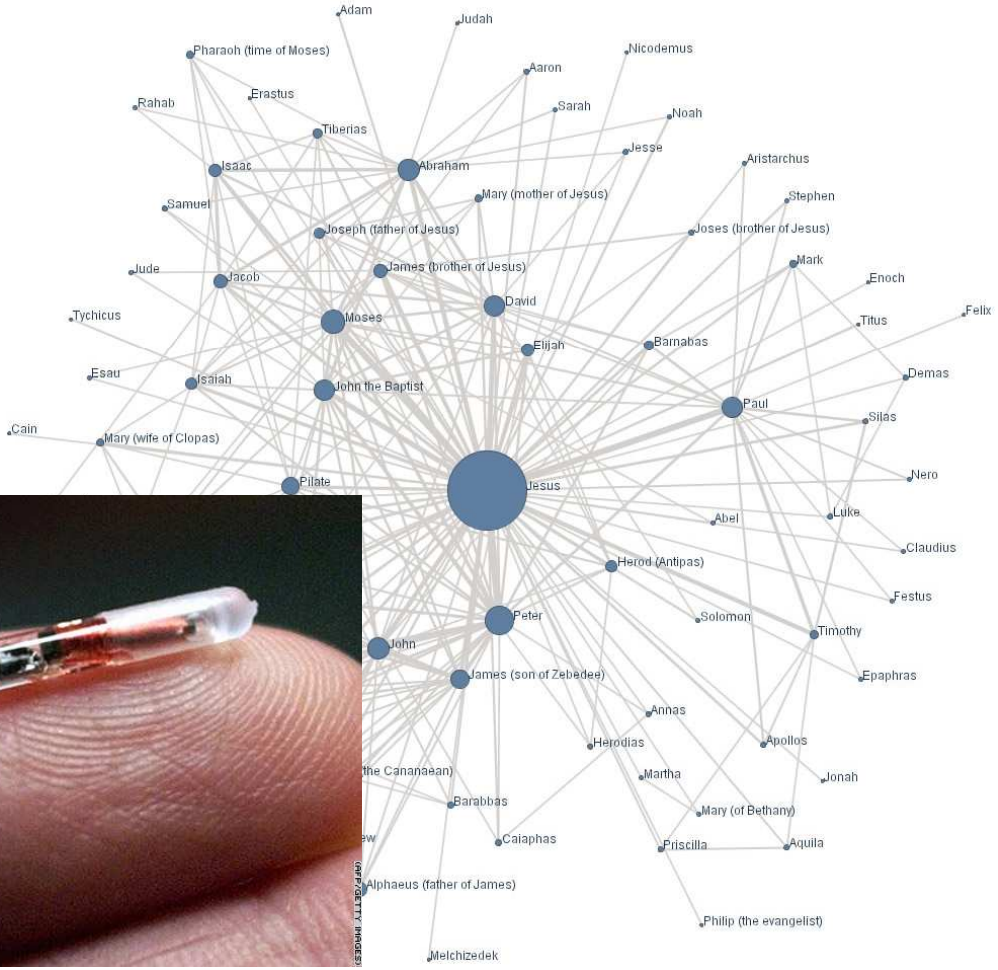
tcrn.ch/nfXTAC

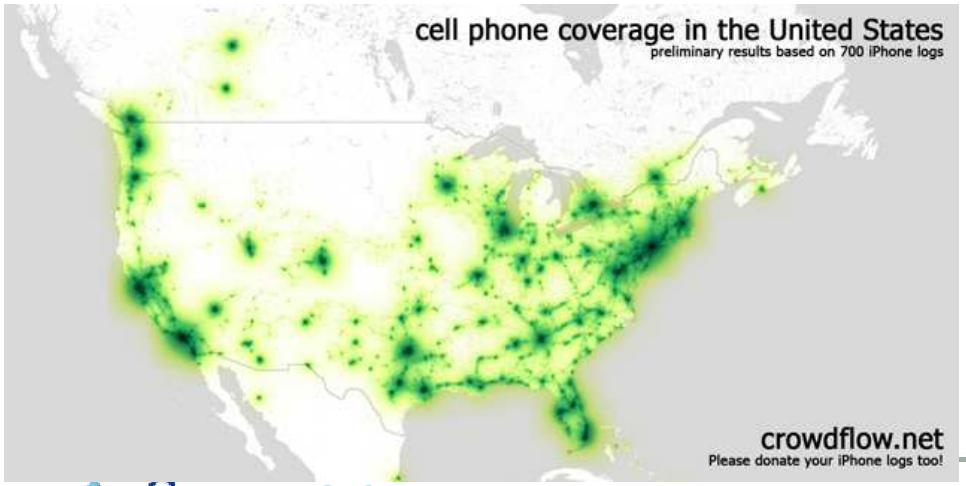
What is BigData ?

2. Variety

- Many more data types (localisation, sensors, etc)







What is BigData ?

3. Velocity

- Need for real-time (or business-time)
- Data lifecycles are accelerating

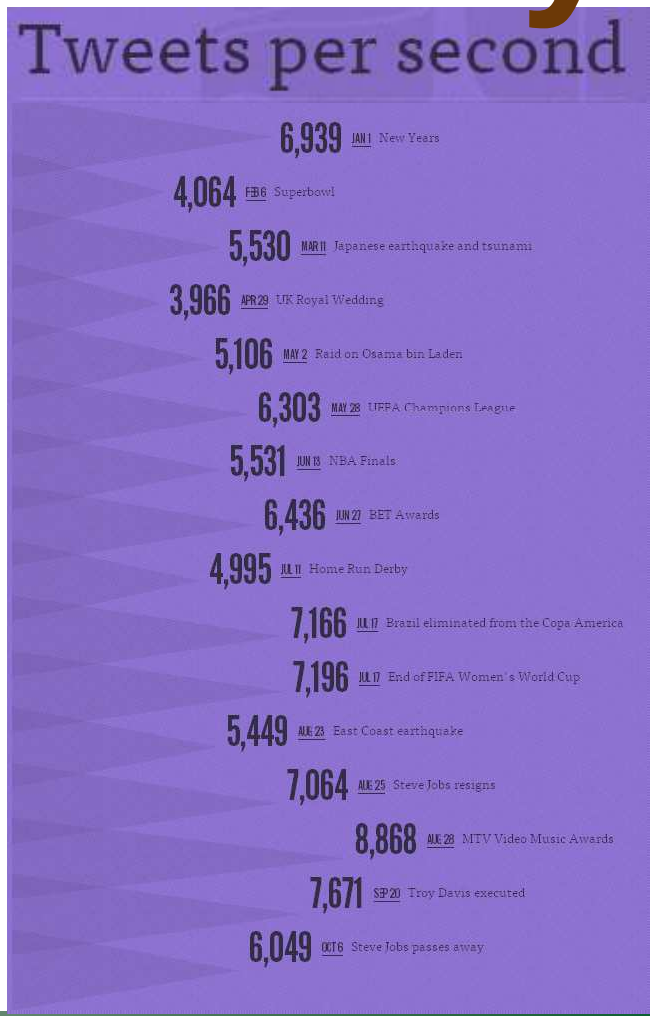
Three of Top Four Highest Tweets-Per-Second Records Set Already in 2012

by DAVE LARSON

"Twitter Comms
@twittercomms

On Dec 9, the television screening in Japan of Hayao Miyazaki's "Castle in the Sky" led to 25,088 Tweets per second - a new Twitter record.

bit.ly/xCT0Mf



What is BigData ?

4. Agility

- Business driven
- End-to-end approach
- Experimental
- Time-to-value over guaranteed success

"Big Data analytics must be business led, and not all projects will be successful at finding the needle in the haystack."

bit.ly/pW9Tj9

What is BigData ?

BigData is not only about storage & volume

**The whole value chain matters,
from a business point of view.**

analytics

visualizing

capture

storage

search

sharing

1. Volume
2. Variety
3. Velocity
4. Agility

Approaches
&
Solutions

More demanding environments

Petabytes

Trillions of records

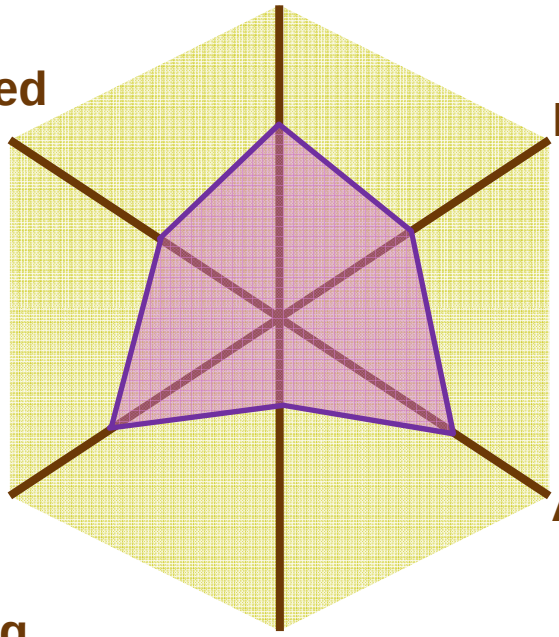
Massive Datasets

Complex relationships

Schema-less

Unstructured Data

Flexible Data model



Scalable I/O & Processing

Advanced Analytics

Trends, Statistics
Predictive modeling
Simulation Models

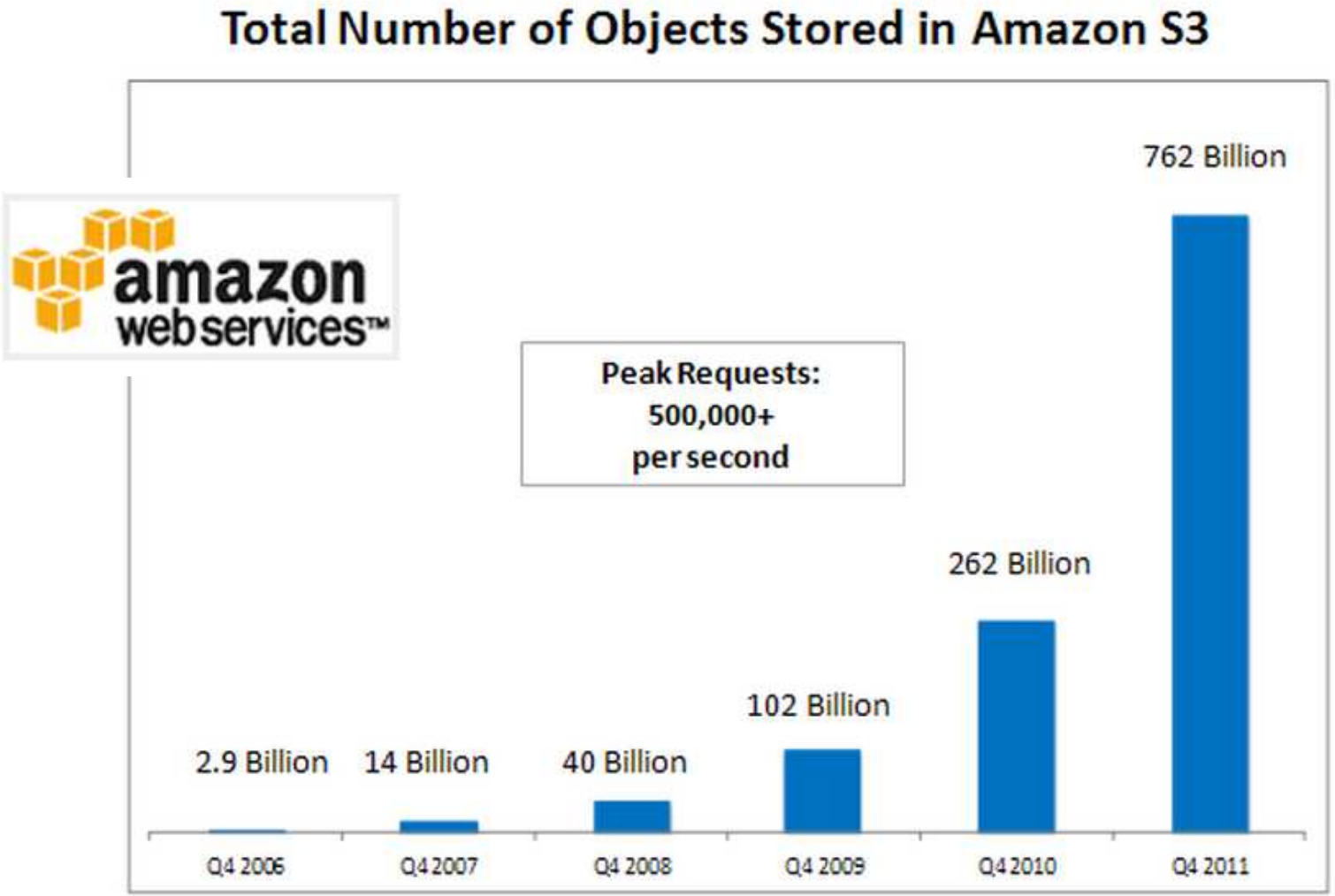
Real-Time

Thousands of concurrent accesses
Continuous data loading

Sub-second Latency

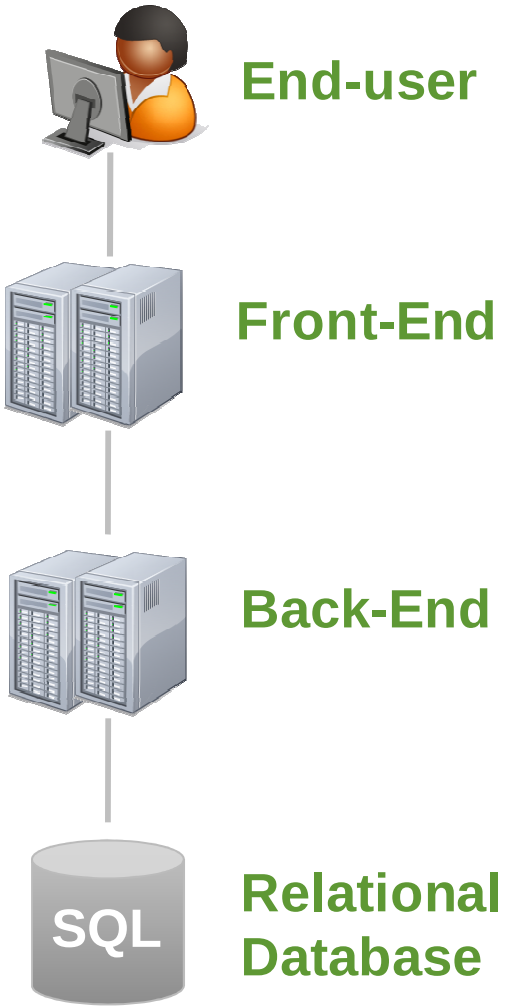
What are your needs ?

Big Data, Cloud & Web-scale platforms are real

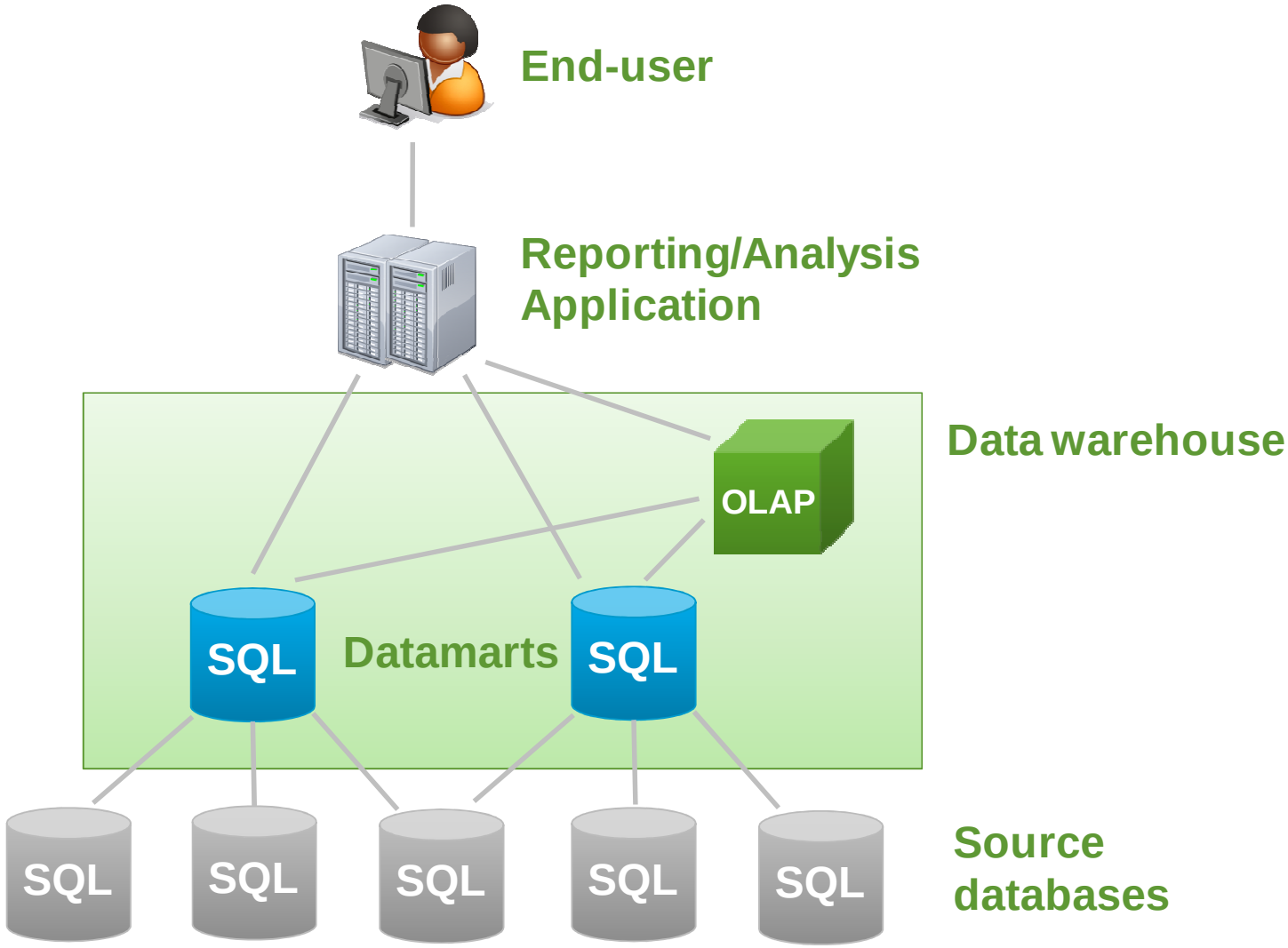


bit.ly/Ah9hjY

Transaction-oriented apps
of the last century



« Classical » BI



SQL

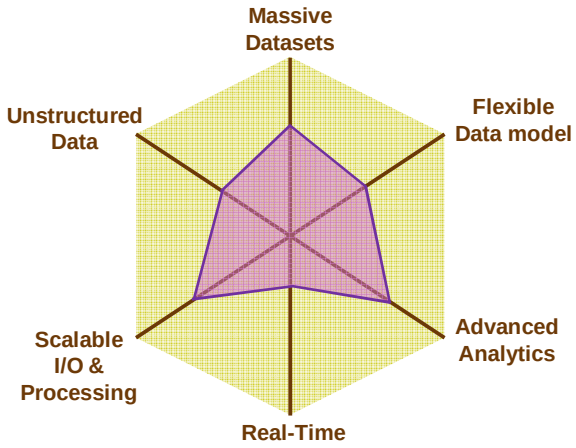
General-purpose.
One size fits all.

NoSQL

Choose the one that best fits…
…according to what you need

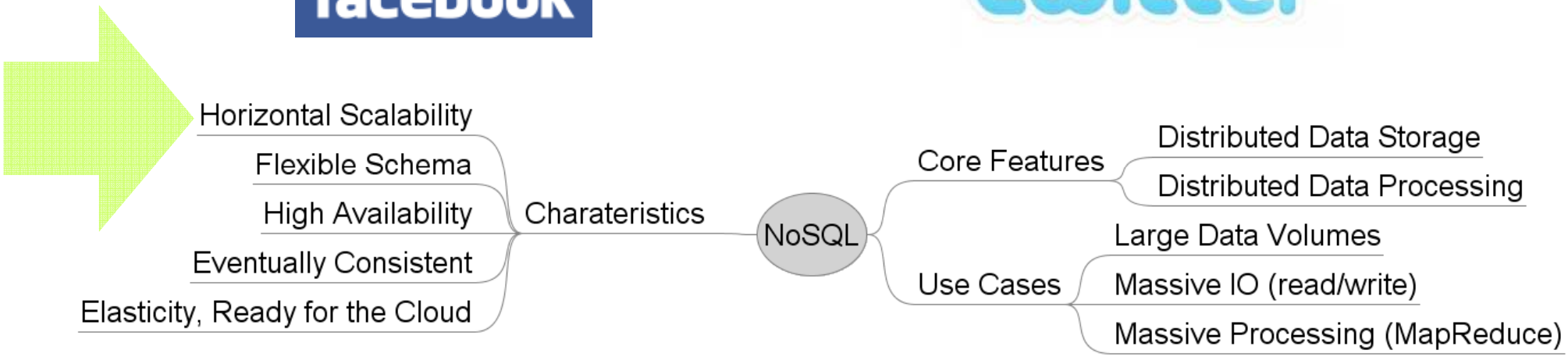
NoSQL = Not only SQL

An familly of data management solutions targeting Big Data requirements



NoSQL: modern web-scale databases

Google LinkedIn YAHOO! facebook twitter



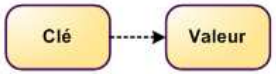
Cassandra

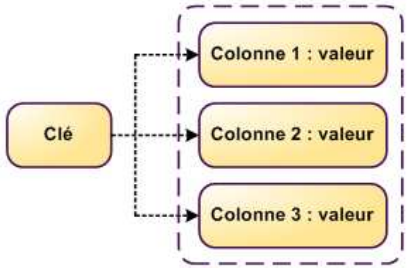
400-node Cassandra architecture for the analysis of hundreds of millions of intelligence documents

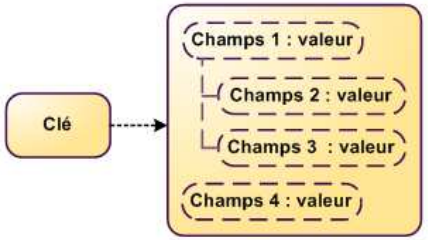
hadoop

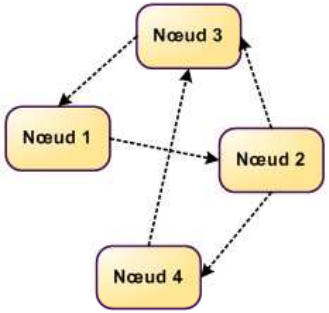
Yahoo! runs Hadoop on 42,000 nodes holding 180-200 petabytes of data

NoSQL: 4 datamodels









Source: Xebia

Key-Value stores

Big Table clones (column family)

Document databases

Graph databases

riak

Cassandra

mongoDB

Neo4j the graph database

NoSQL: additional charateristics

A whole ecosystem.
Many kinds of approaches & products.

	Data Model	Query API	Persistence Design
Cassandra	Columnfamily	Thrift	Memtable / SSTable
CouchDB	Document	map/reduce views	Append-only B-tree
HBase	Columnfamily	Thrift, REST	Memtable / SSTable on HDFS
MongoDB	Document	Cursor	B-tree
Neo4J	Graph	Graph	On-disk linked lists
Redis	Collection	Collection	In-memory with background snapshots
Riak	Document	Nested hashes	?
Scalaris	Key/value	get/put	In-memory only
Tokyo Cabinet	Key/value	get/put	Hash or B-tree
Voldemort	Key/value	get/put	Pluggable (primarily BDB MySQL)

http://www.rackspace.com/cloud/blog/2009/11/09/nosql-ecosystem/

Multiple approaches & solutions

SBAs
Search-Based Applications

exalead

kaggle™

Crowdsourcing

VERTICA An HP Company

Columnar databases

Batch-Oriented / Stream-Oriented

Self-service BI

In-database processing

NoSQL

hadoop

Data-Grid

In-Memory

Appliances (HW+SW)

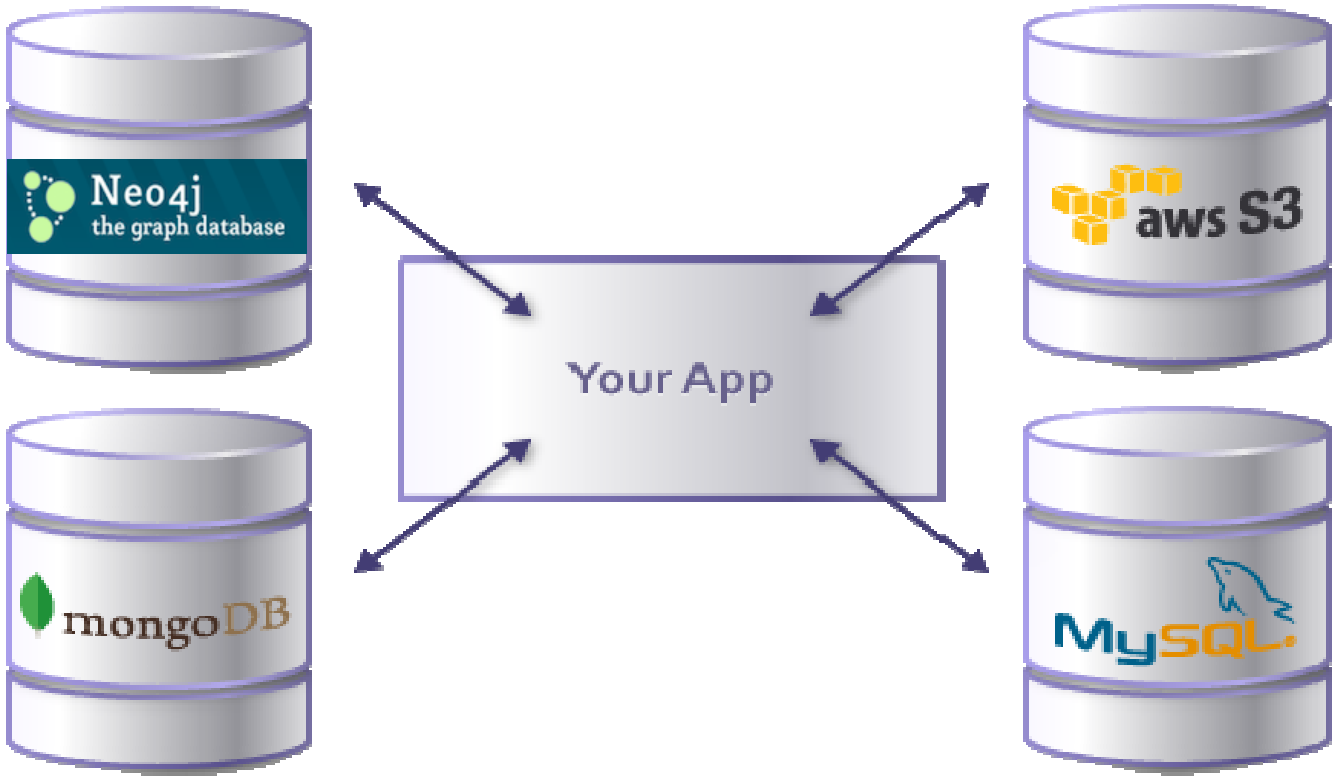
SAP® IN-MEMORY APPLIANCE (SAP HANA™)
THE NEXT WAVE OF SAP® IN-MEMORY COMPUTING TECHNOLOGY

IBM WATSON

ORACLE
BIG DATA APPLIANCE

Oracle NoSQL Database
Cloudera Manager
Cloudera CDH
Oracle Big Data Connectors
Oracle Enterprise Linux & Java VM
Oracle R

Hybrid Architectures – Polyglot Persistence



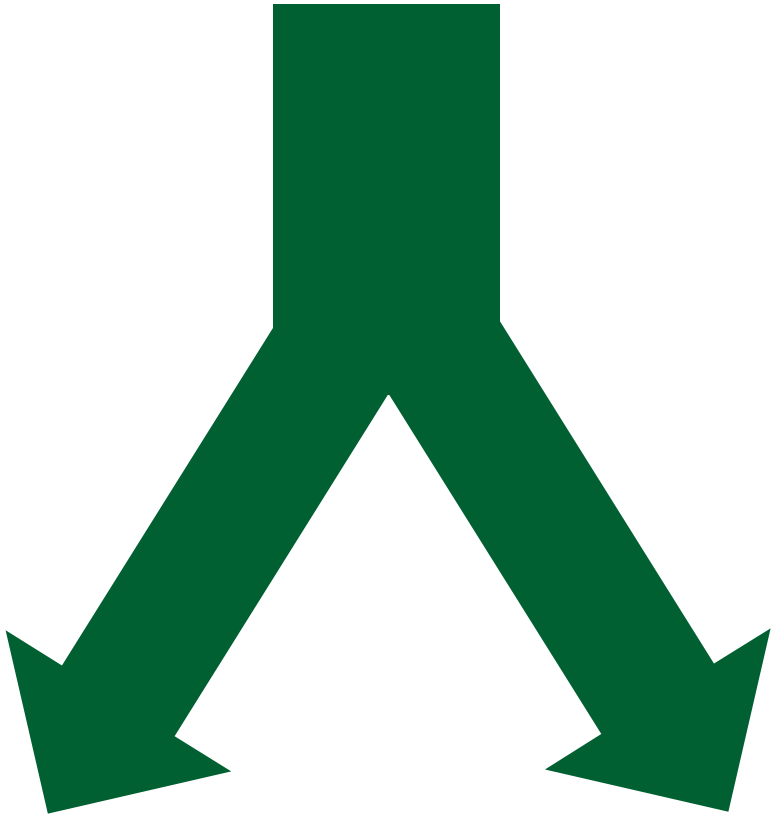
Selecting one single solution is not the target !

Separate MDM & Big Data approaches

ACID
Atomicity, Consistency, Isolation, Durability

Governance
Data Quality
Normalized

Data as critical asset



BASE
Basically Available, Soft state, Eventually consistent

Agility
Scalability
Denormalized

Data as raw material for analysis & insight

MDM
(Master Data Management)

Big Data

Customer Case

Customer Case: Aircrafts Data

YouTube

90

Flight hours

~160 MBytes

48

Hours of video uploaded

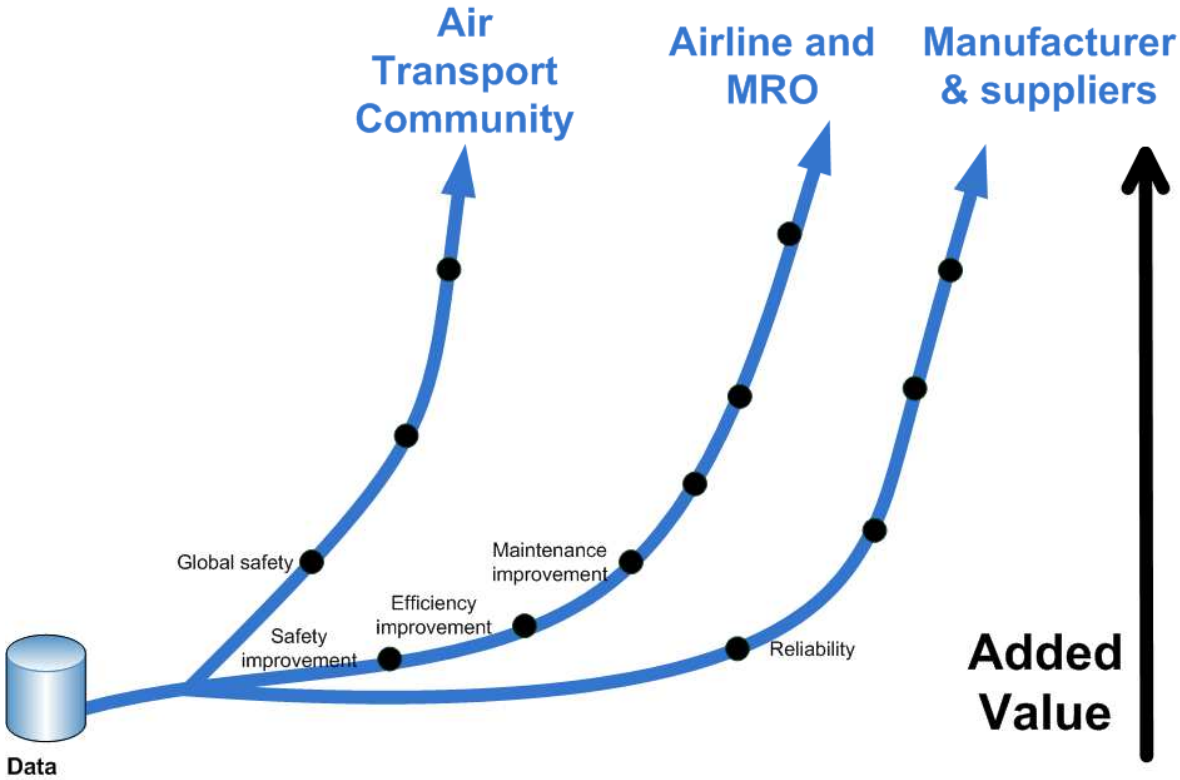
THINGS THAT HAPPEN ON INTERNET EVERY SIXTY SECONDS

(Aircrafts worldwide operations 2010
sources: http://www.boeing.com/news/techissues/pdf/statsum.pdf
http://www.youtube.com/t/press_statistics)

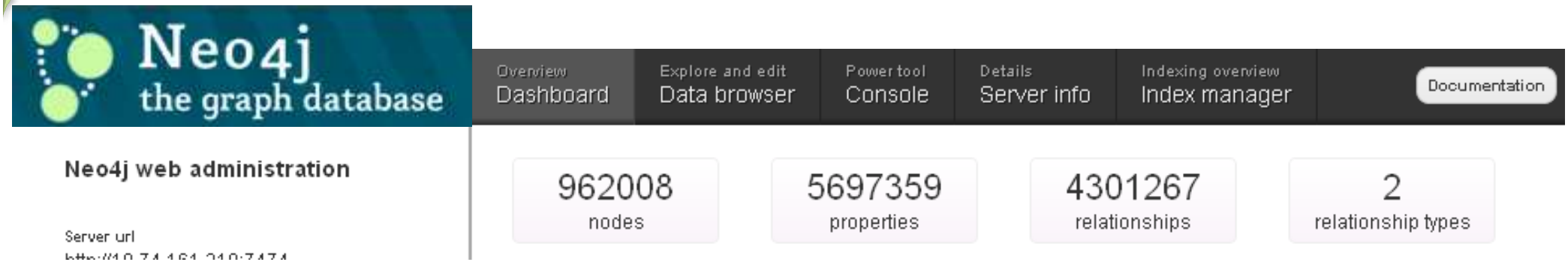
bit.ly/ySQR7g

Customer Case: Big Data ➔ Big Value

- Leverage Aircraft Data…
- … to enable new value-added services…
- … delivered in a new way

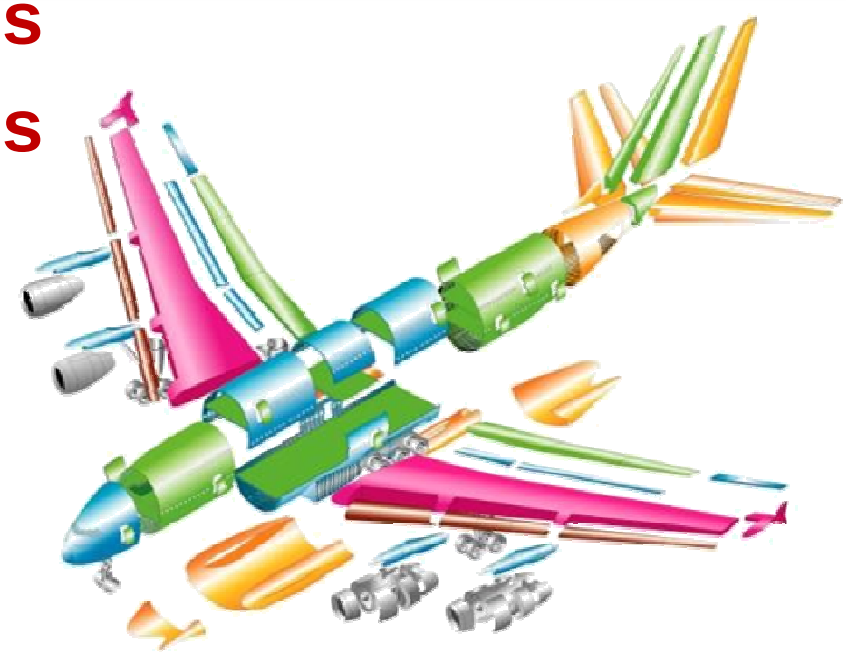


Customer Case: Aircraft Product Structure



- Check Data Quality **~100 ms**
- Traverse product structure **~400 ms**
  (extract all DS for an effectivity)
- Manufacturing use case **~7 s**
  (complete BOM for an effectivity)

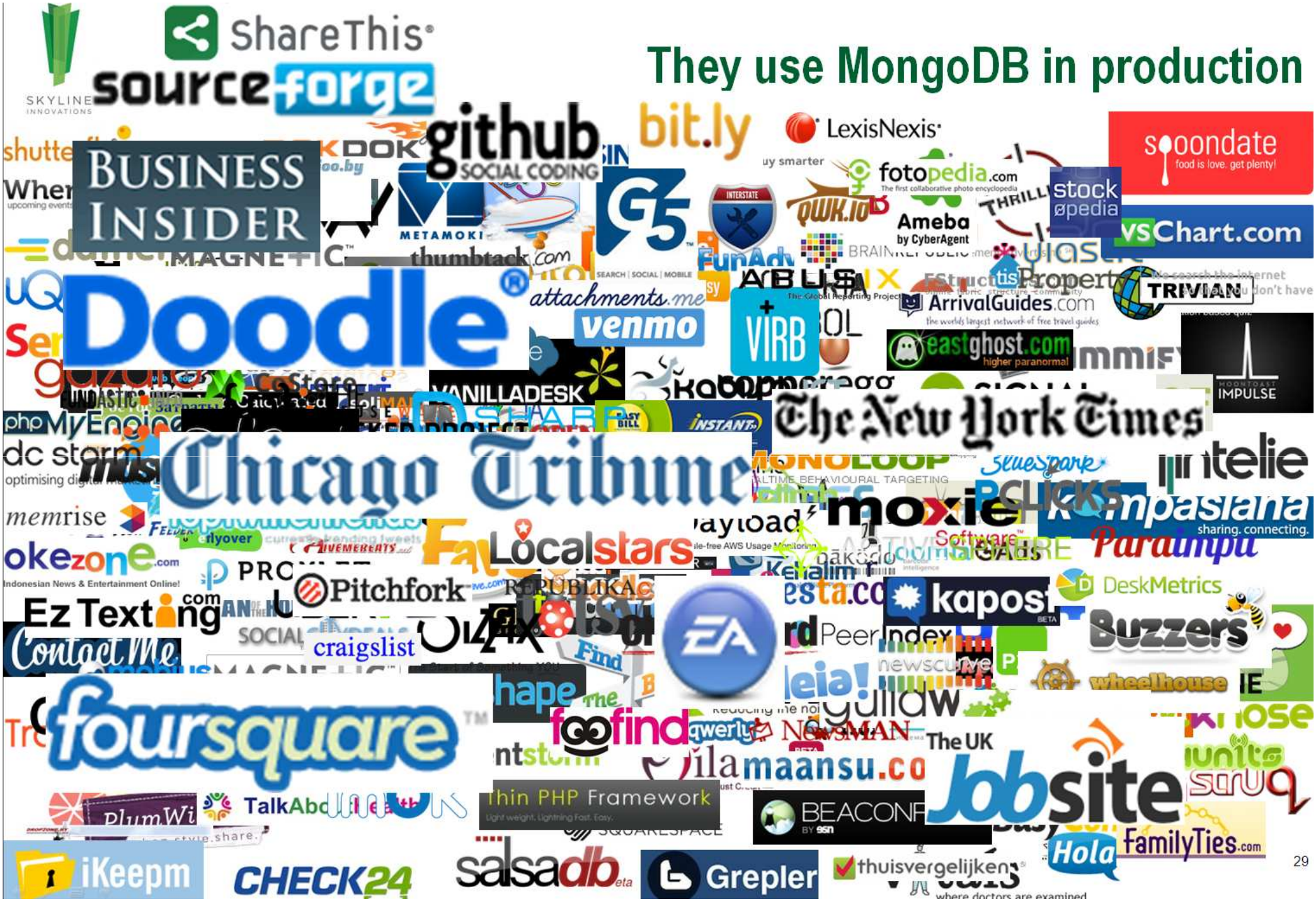
~ 1 hour
with current solution (RDBMS)



Conclusion
Principles & Trends

- Big Data ➔ Big Opportunity
- An end-to-end approach implying a new mindset
  - IO, Storage, Query, Analysis all together
- High Diversity & Specialization of the available solutions
  - All major vendors: **HP** (acquired Vertica), **IBM** (acquired Netezza), **SAP**, …
  - Many startups: **1**0gen, **A**cunu, **B**asho, **C**loudera, **D**atastax… **Z**illabyte
  - Get ready to use several solutions
- Choose the right tool for the right job…
- …but do not take too much time to think

They use MongoDB in production



Thank you !

olivier.flebus@capgemini.com

@olivierflebus

Meet our experts
Benchmark with other customers

http://www.capgemini.com/architectureweek/

Capgemini

CONSULTING.TECHNOLOGY.OUTSOURCING

www.capgemini.com