



Couchbase



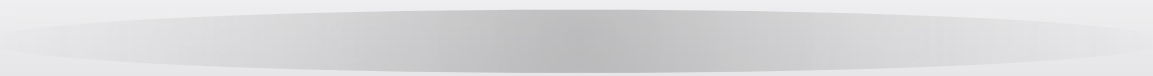
How to integrate Hadoop with your NoSQL database?

Tugdual “Tug” Grall

Technical Evangelist



Couchbase



About Me

- **Tugdual “Tug” Grall**

- Couchbase
 - Technical Evangelist
- eXo
 - CTO
- Oracle
 - Developer/Product Manager
 - Mainly Java/SOA
- Developer in consulting firms

- Web

 @tgrall

 <http://blog.grallandco.com>

 tgrall

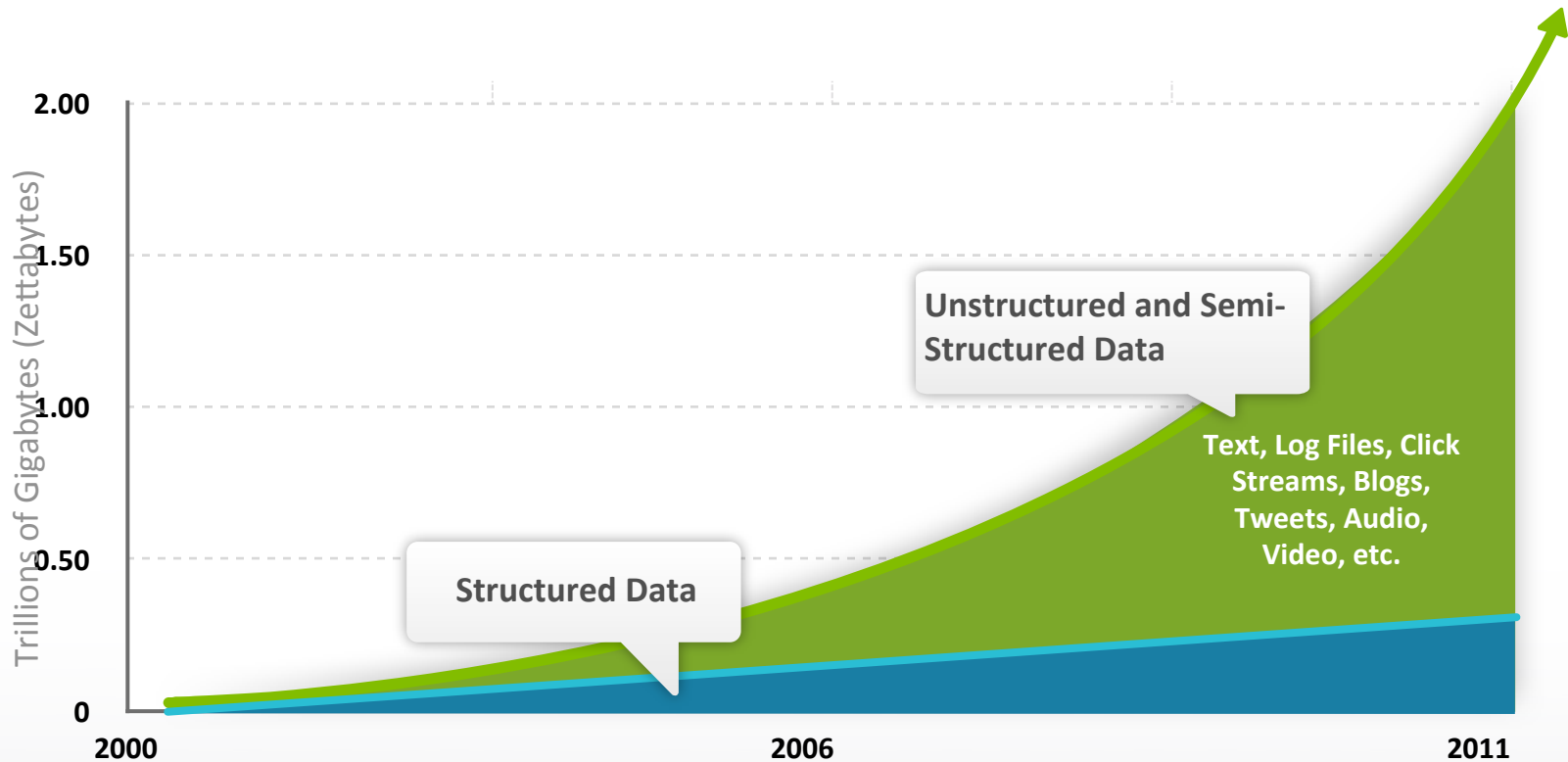
- NantesJUG co-founder

- Pet Project :

- <http://www.resultri.com>

Big Data

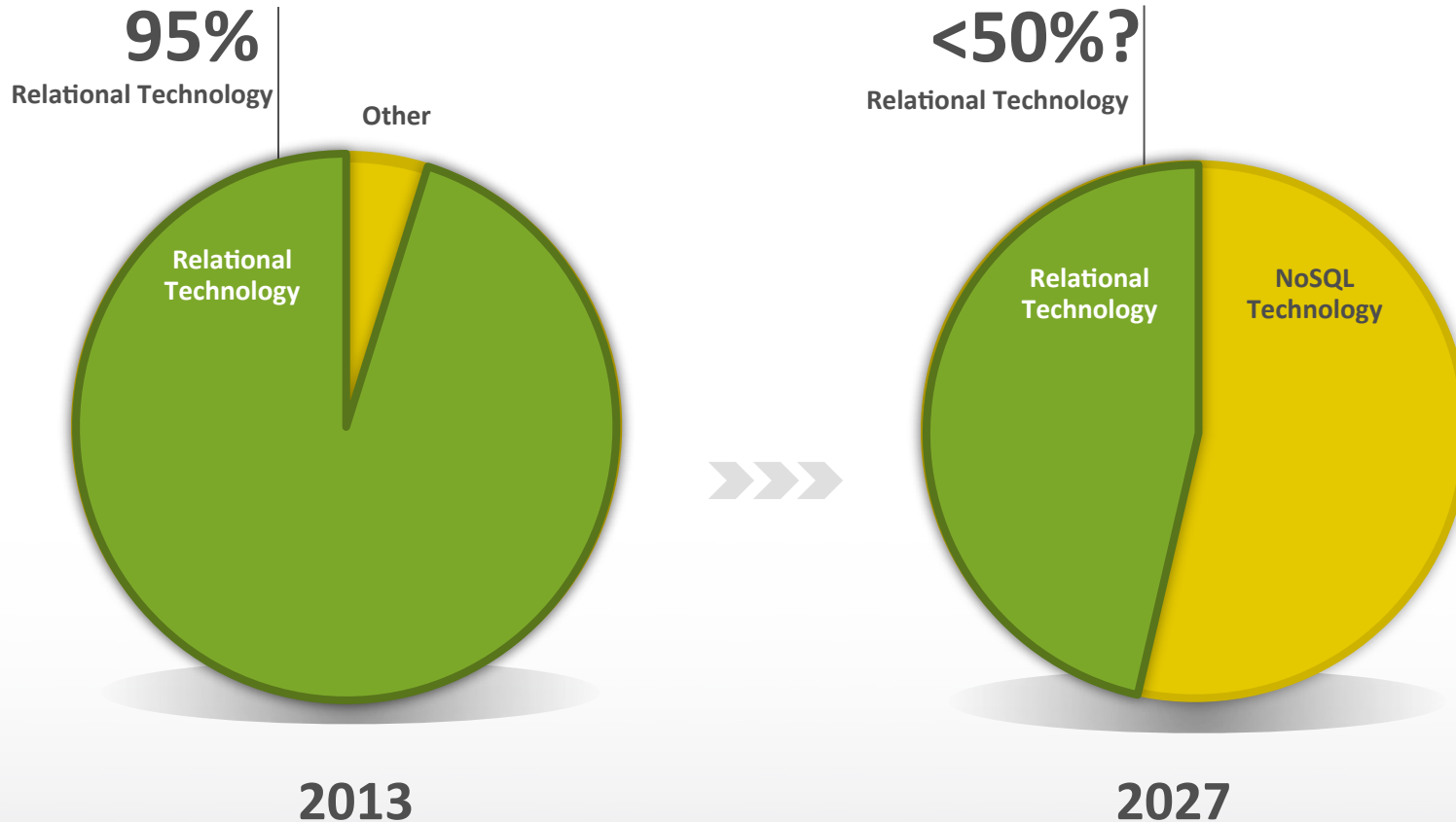
High Data Variety and Velocity



Source: IDC 2011 Digital Universe Study (<http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>)

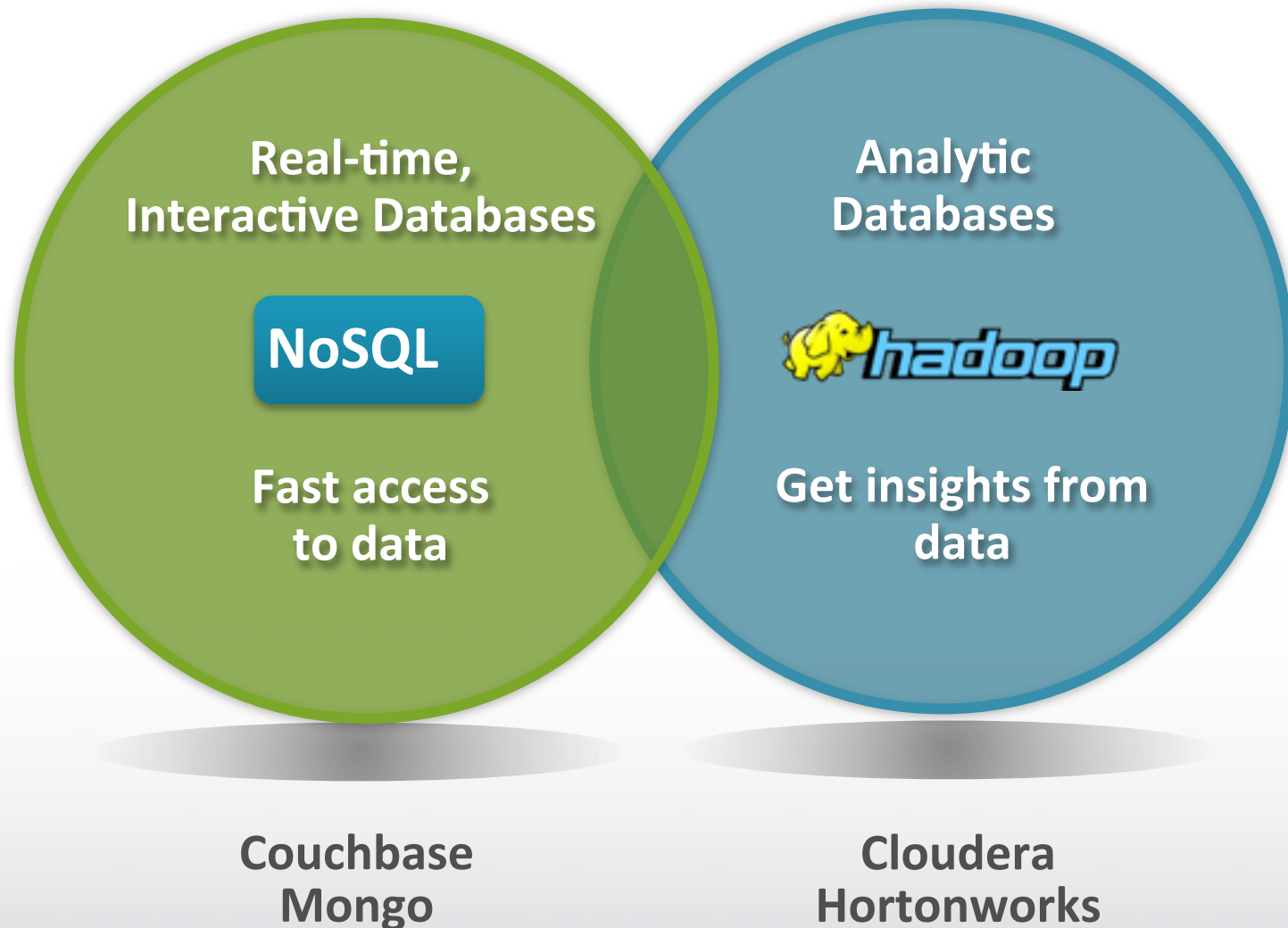
More Flexible Data Model Required

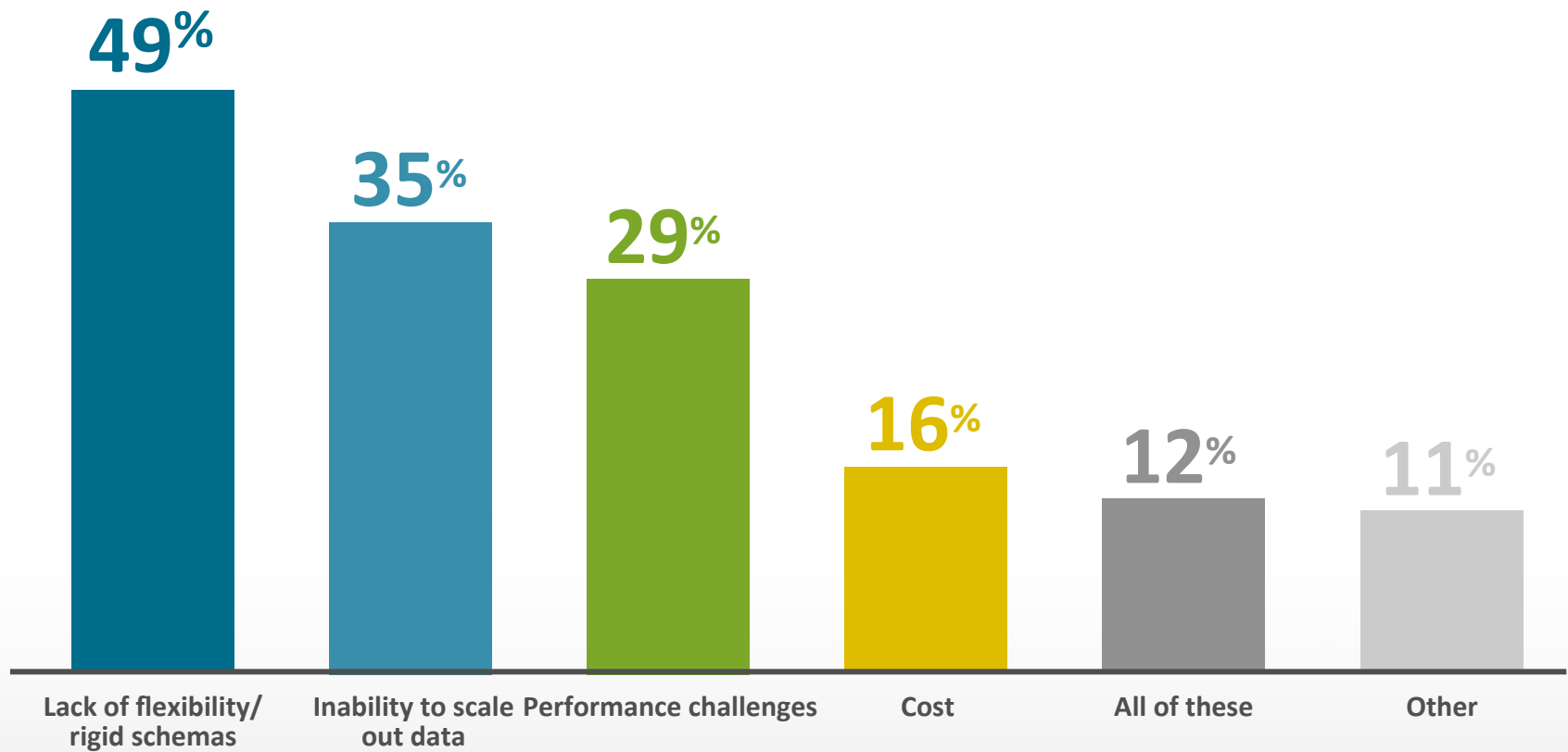
\$30B Database Market Being Disrupted



All new database growth will be NoSQL

Operational vs. Analytic Databases





Source: Couchbase Survey, December 2011, n = 1351.



Hadoop

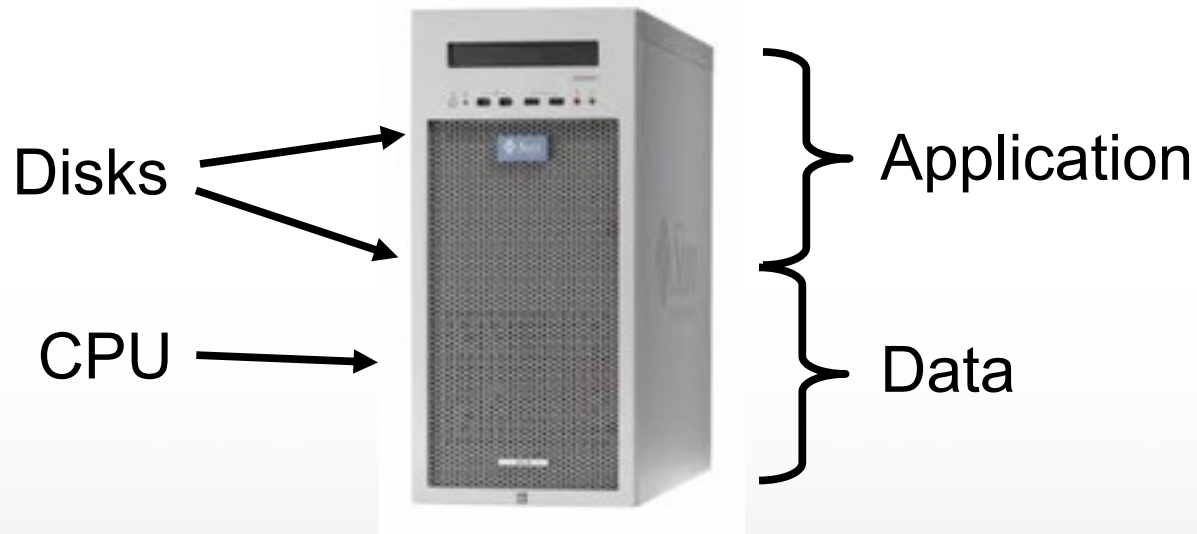


What is Hadoop?

- **Highly scalable**
- **Unstructured data**
- **Open source**
- **Big Data Operating System**
- **Changing the World One Petabyte at a Time**

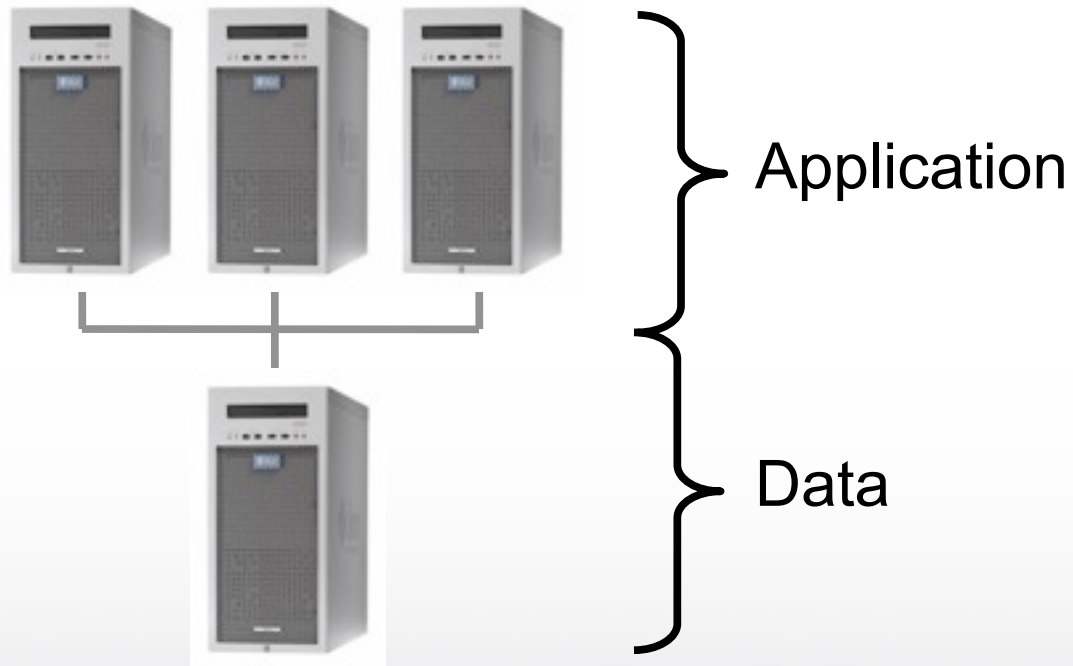
What is Hadoop?

- Simplest unit of compute and storage



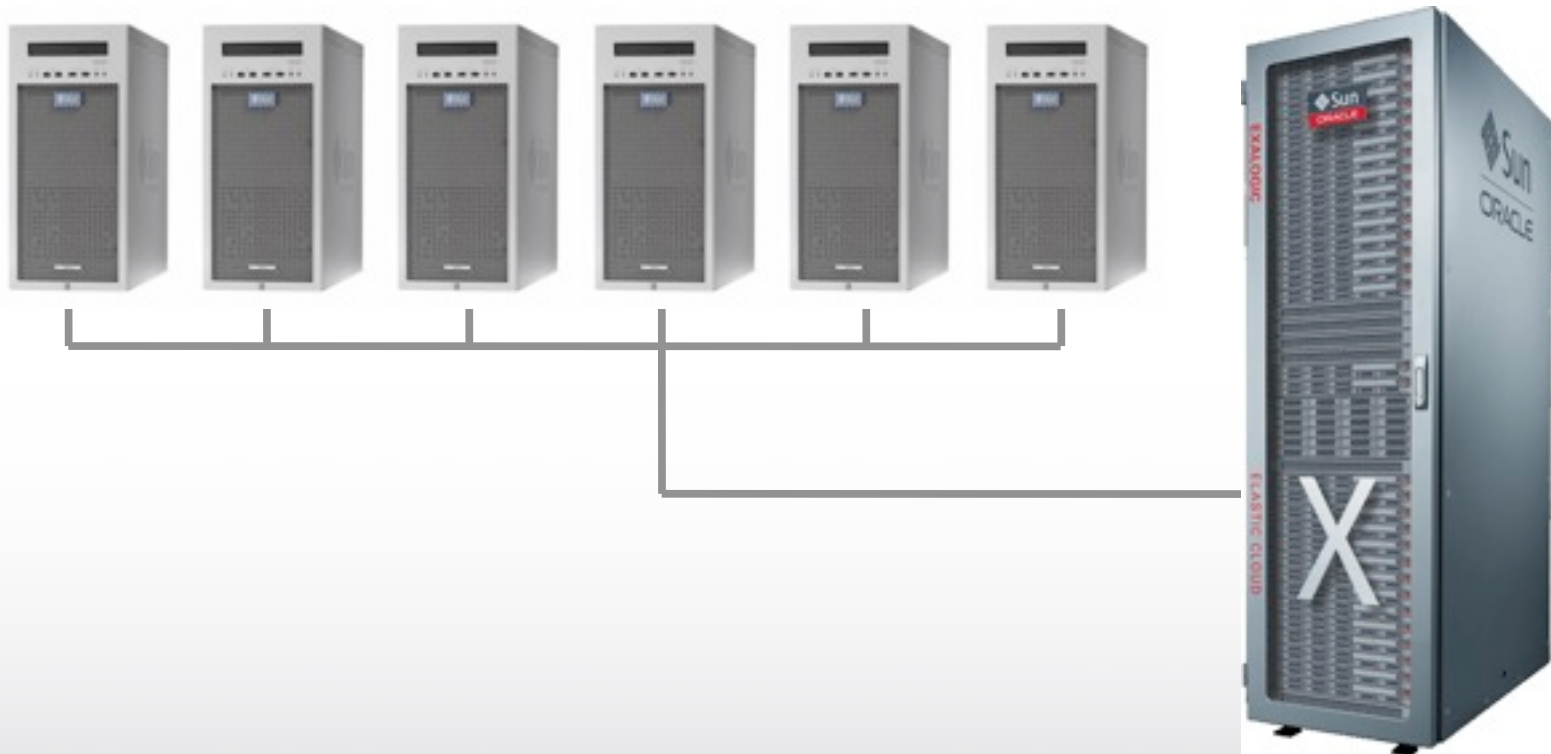
What is Hadoop?

- And when it grows?



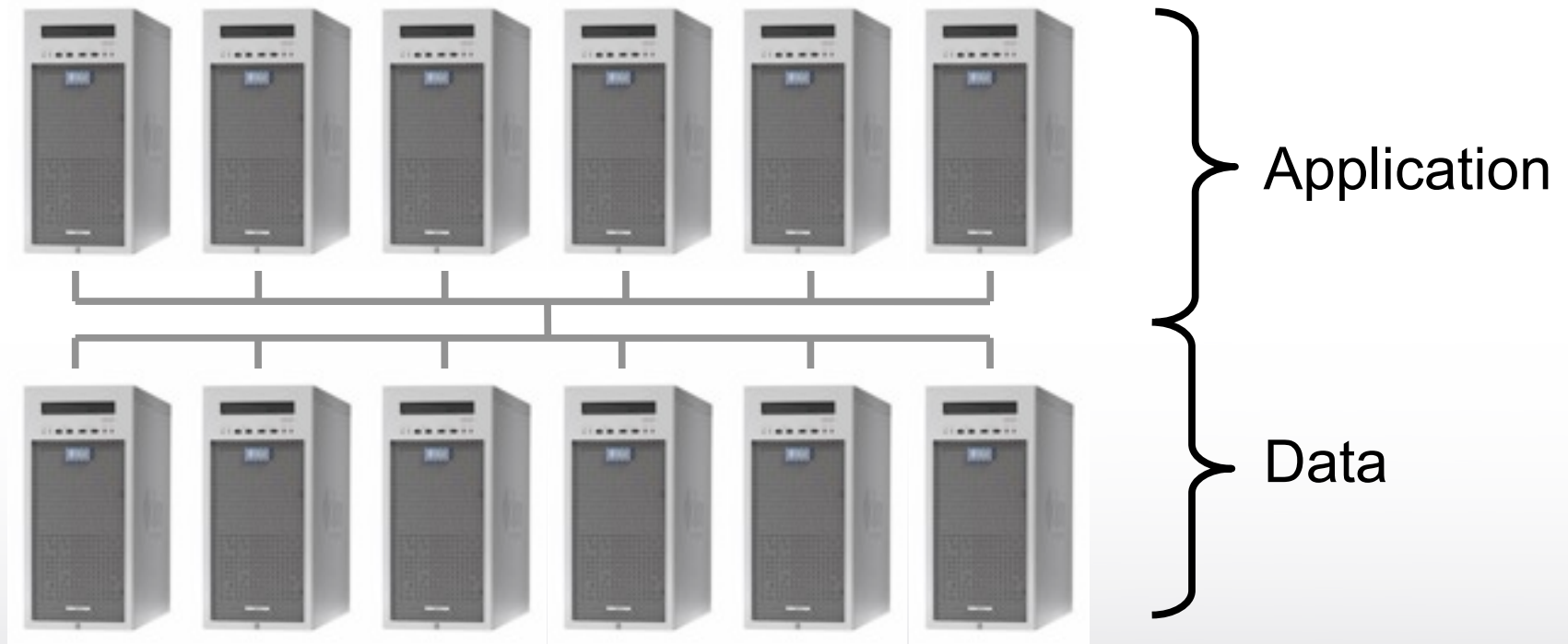
What is Hadoop?

- And when it grows more?



What is Hadoop?

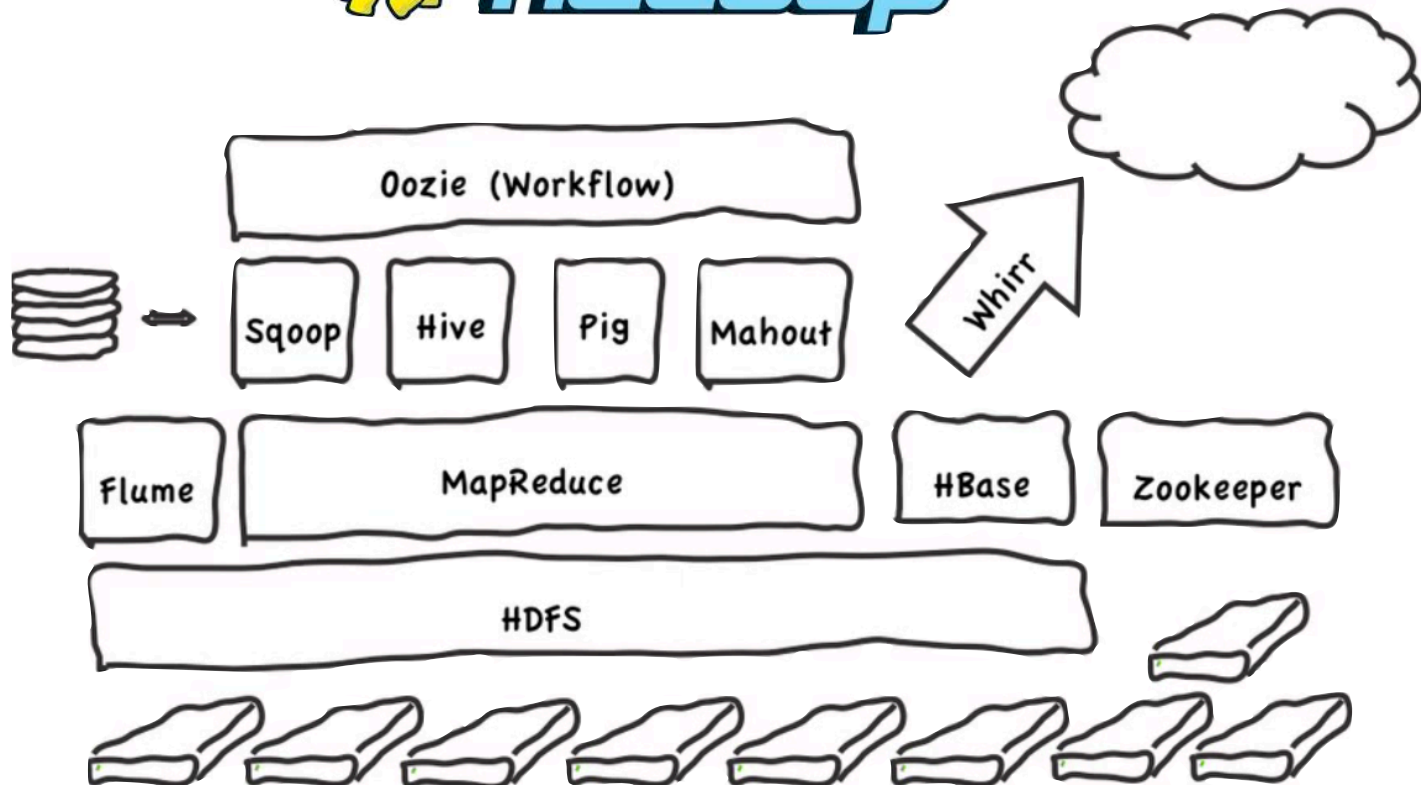
- NoSQL to the rescue



What is Hadoop?

- Hadoop is a different paradigm



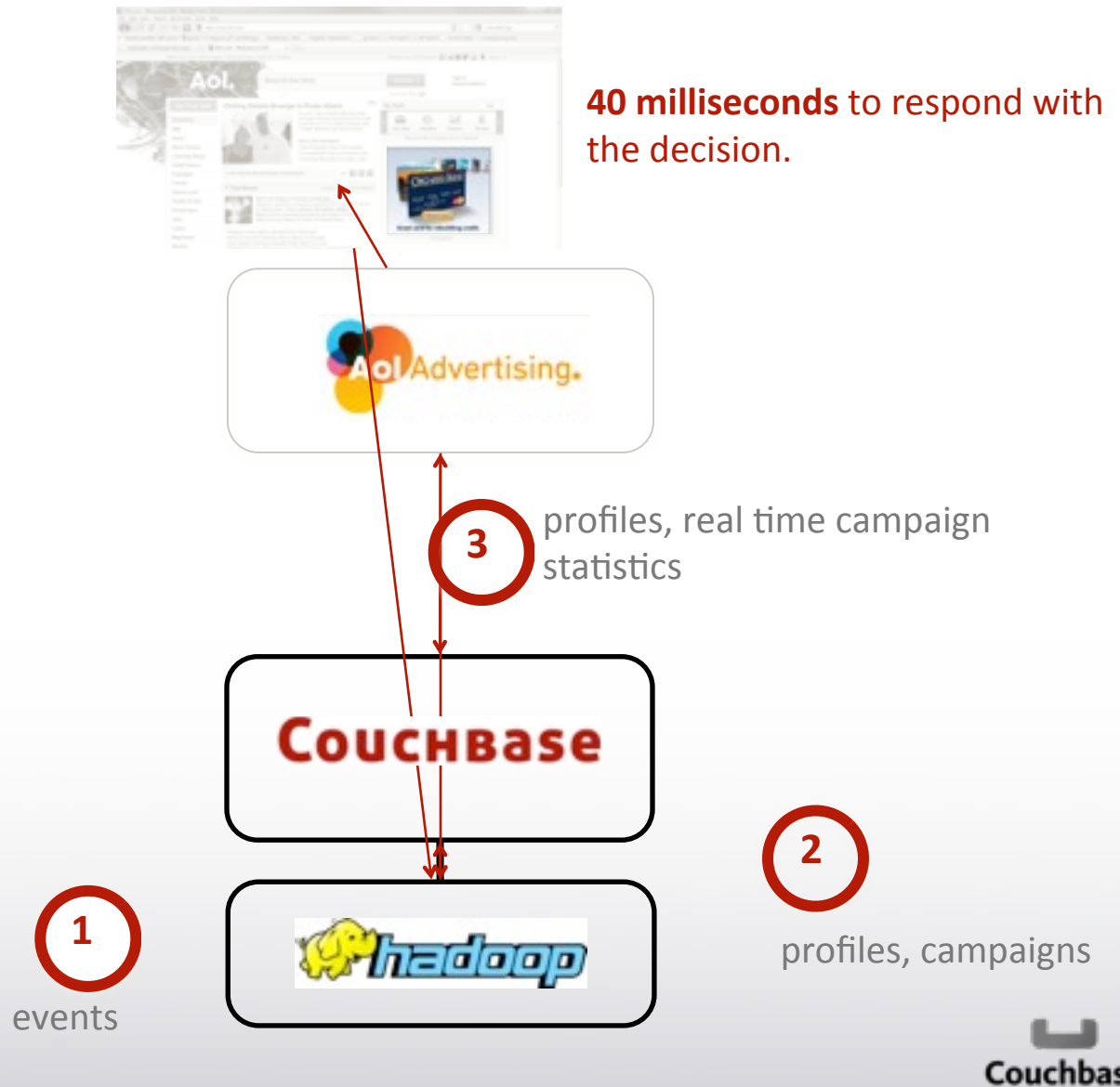




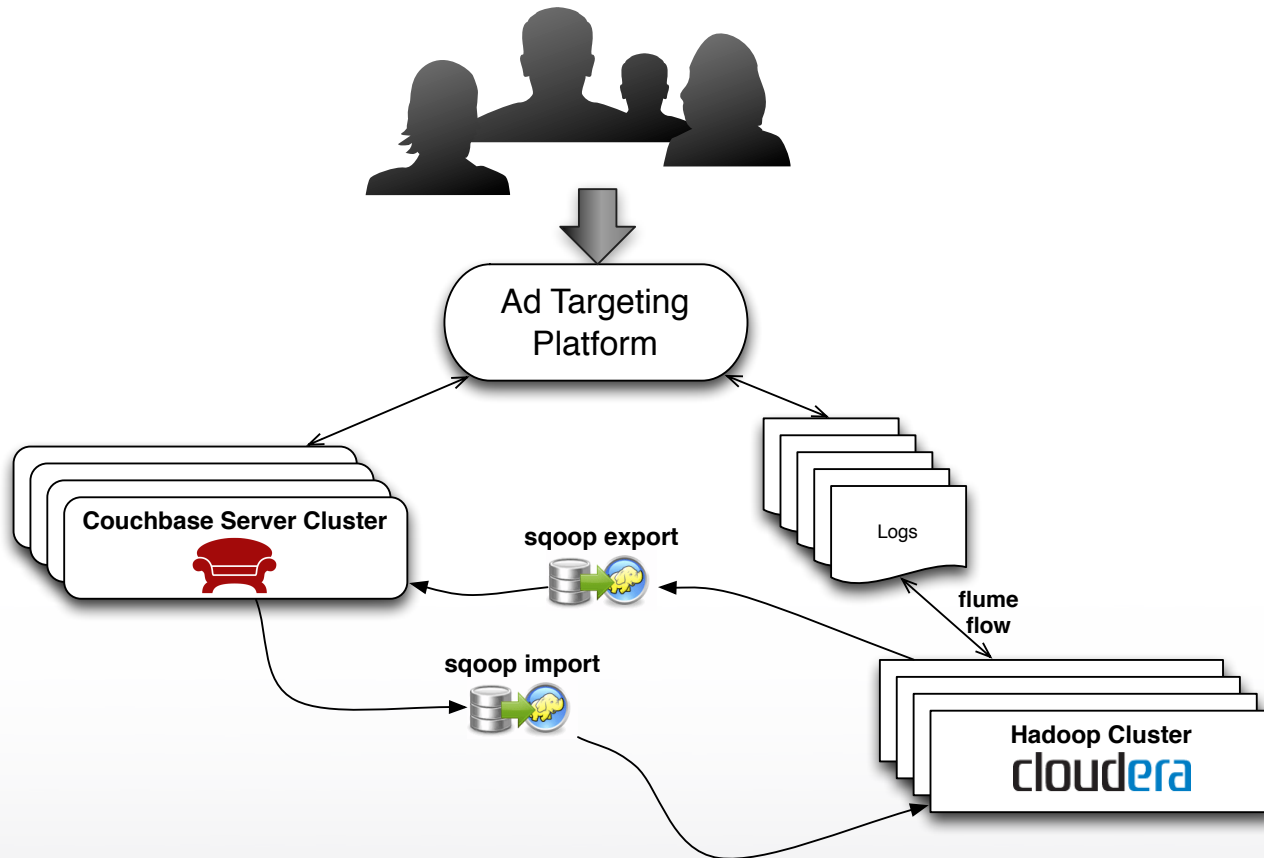
Hadoop and NoSQL



Ad and offer targeting



Moving Parts



Content & Recommendation Targeting

the knot

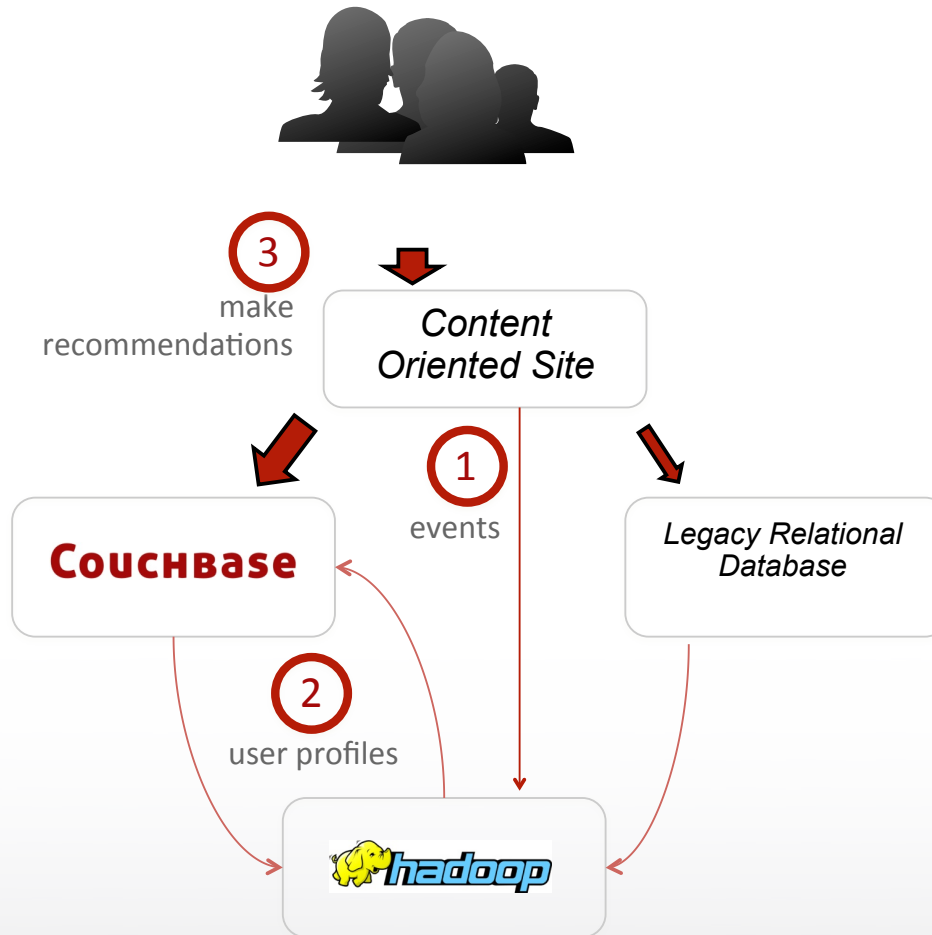
NAVTEQ

|>|<|>|<

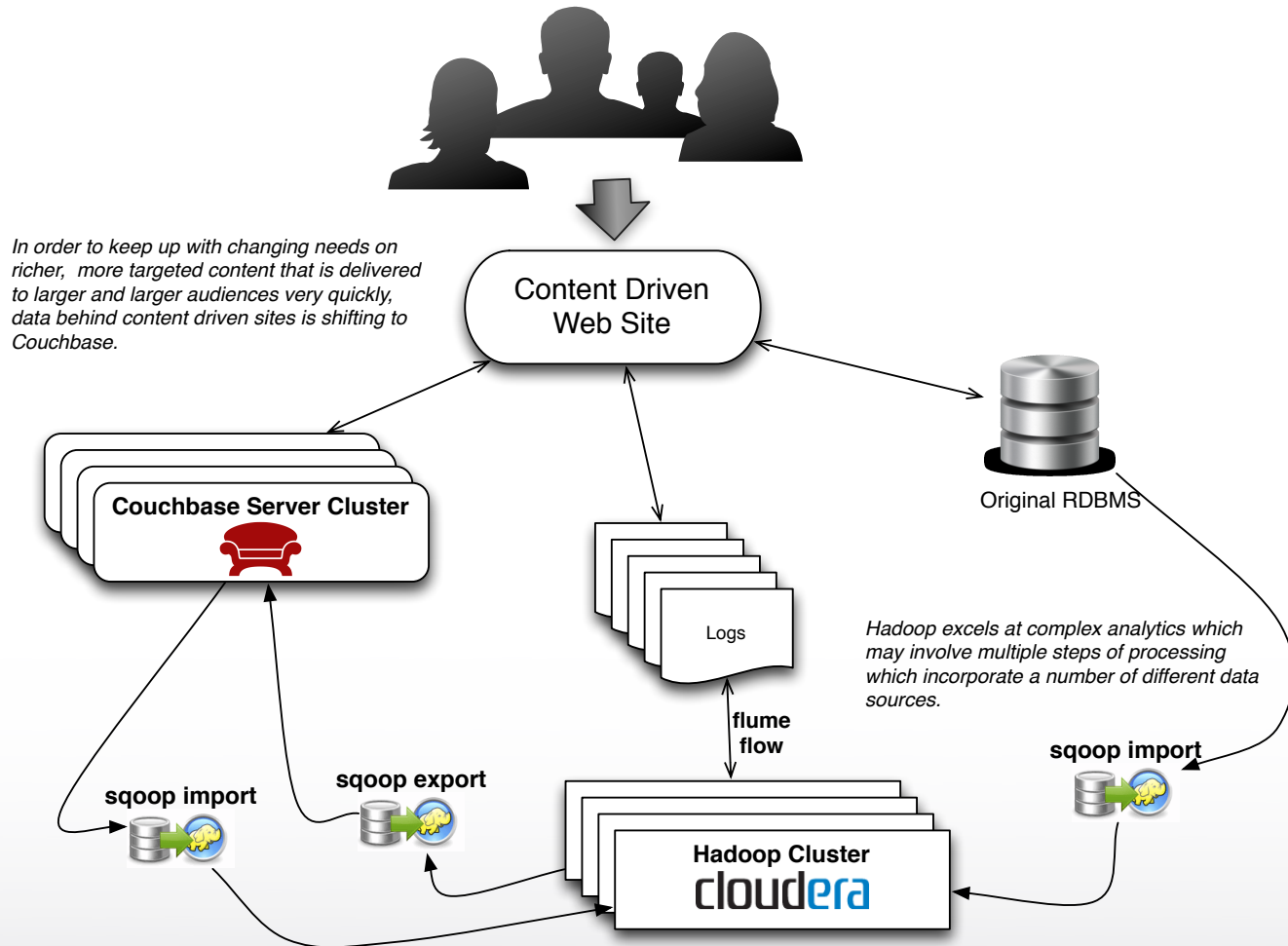
vimeo

mozilla
FOUNDATION

salesforce



Moving Parts



What is Sqoop?

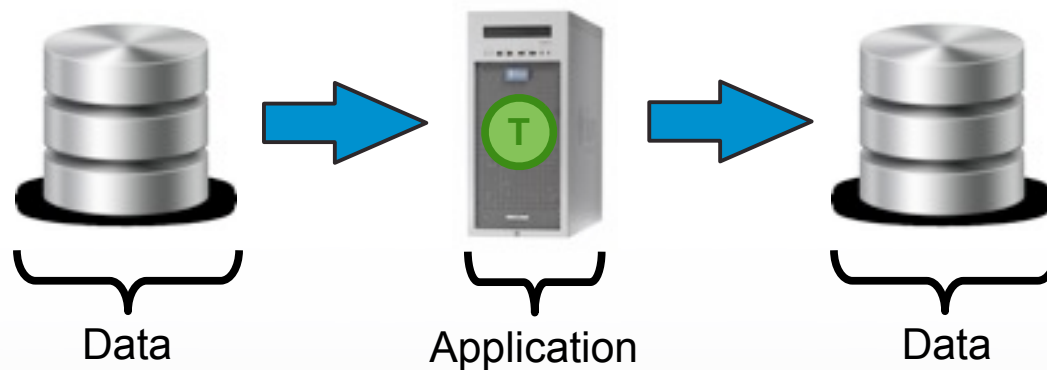
*Sqoop is a tool designed to **transfer data between Hadoop and relational databases**.*

*You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), **transform the data in Hadoop MapReduce, and then export the data back into an RDBMS**.*

sqoop.apache.org

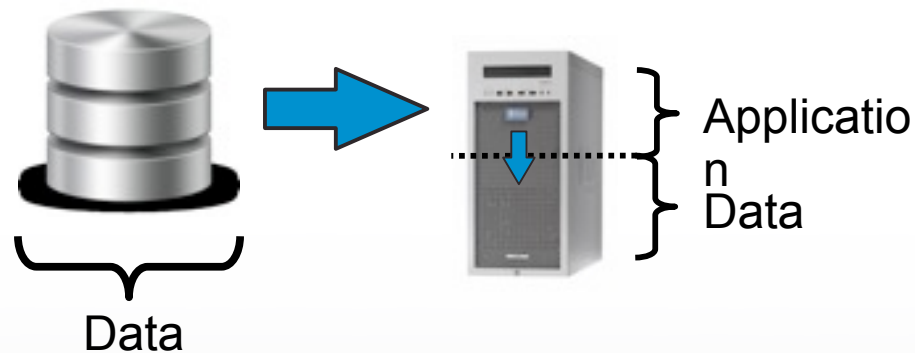
What is Sqoop?

- Traditional ETL



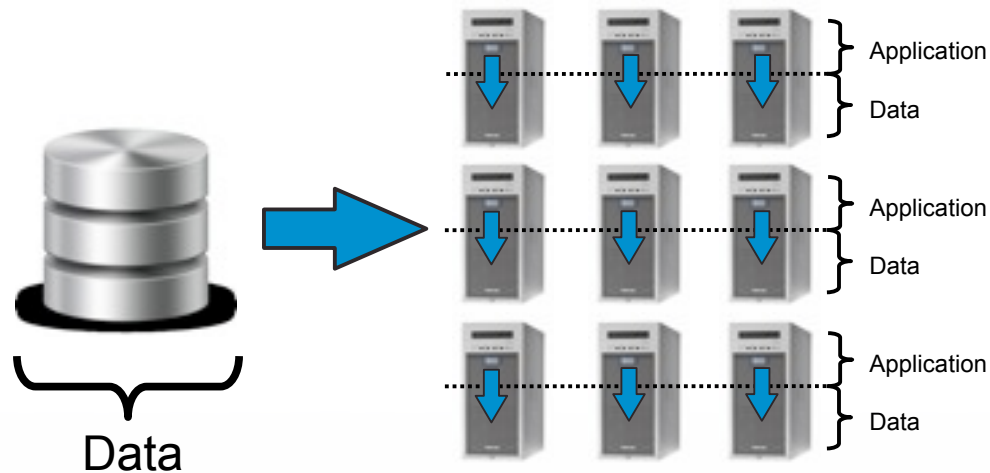
What is Sqoop?

- A different paradigm



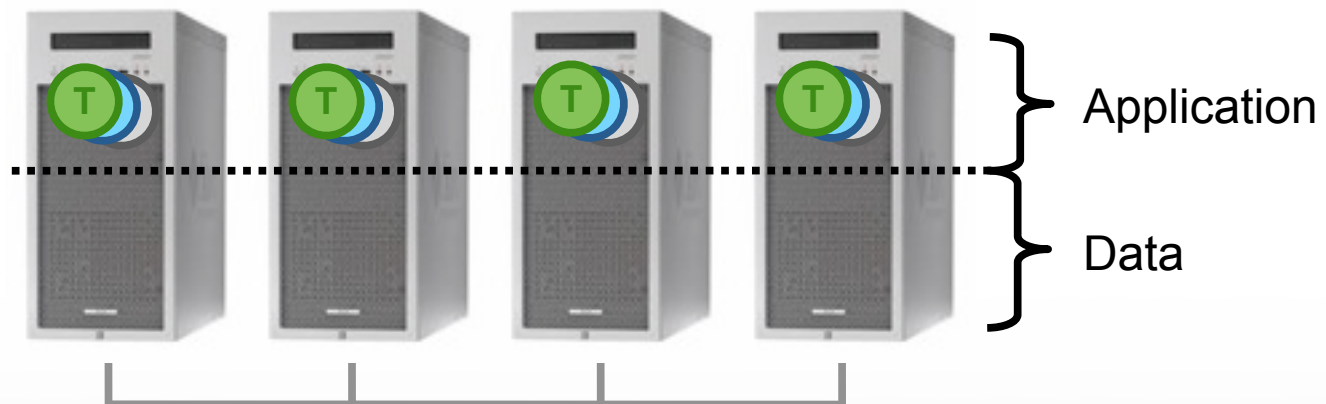
What is Sqoop?

- A very scalable different paradigm



What is Sqoop?

- Where did the Transform go?



What is Sqoop?

- **Sqoop “SQL-Hadoop”**
 - Default connection is via JDBC
- **Lots of custom connectors**
 - Couchbase, VoltDB, Vertica
 - Teradata, Netezza
 - Oracle, MySQL, Postgres

Sqoop : Import



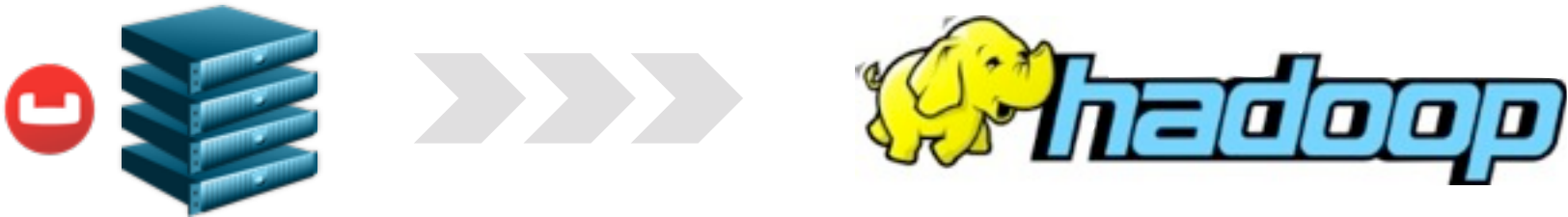
```
sqoop import --connect jdbc:mysql://rdbms1.demo.com/CRM  
--table customers
```

Sqoop : Export



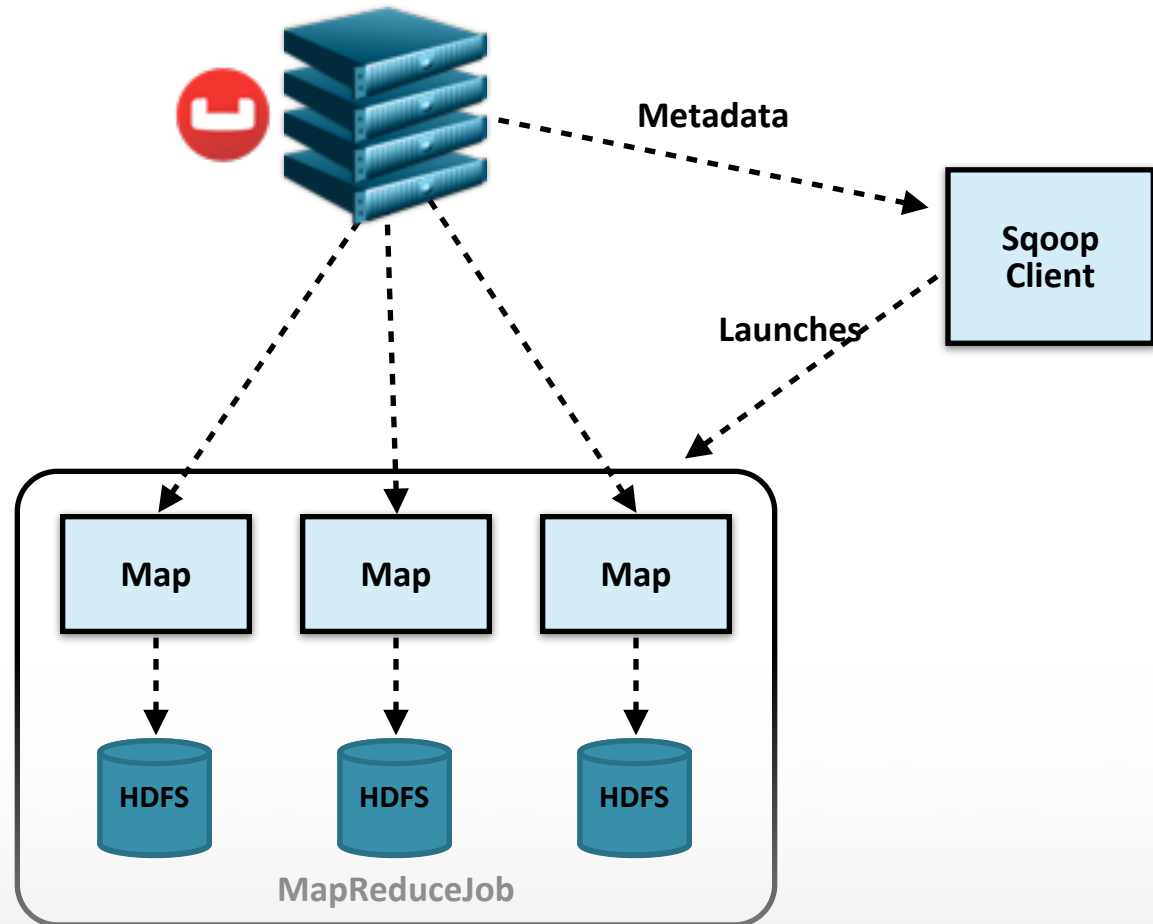
```
sqoop export --connect jdbc:mysql://rdbms1.demo.com/ANALYTICS  
--table sales  
--export-dir /user/hive/warehouse/zip_profits  
--input-fields-terminated-by '\001'
```

Sqoop : Import



```
sqoop import --connect http://localhost:8091/pools  
--table DUMP
```

Sqoop : Import

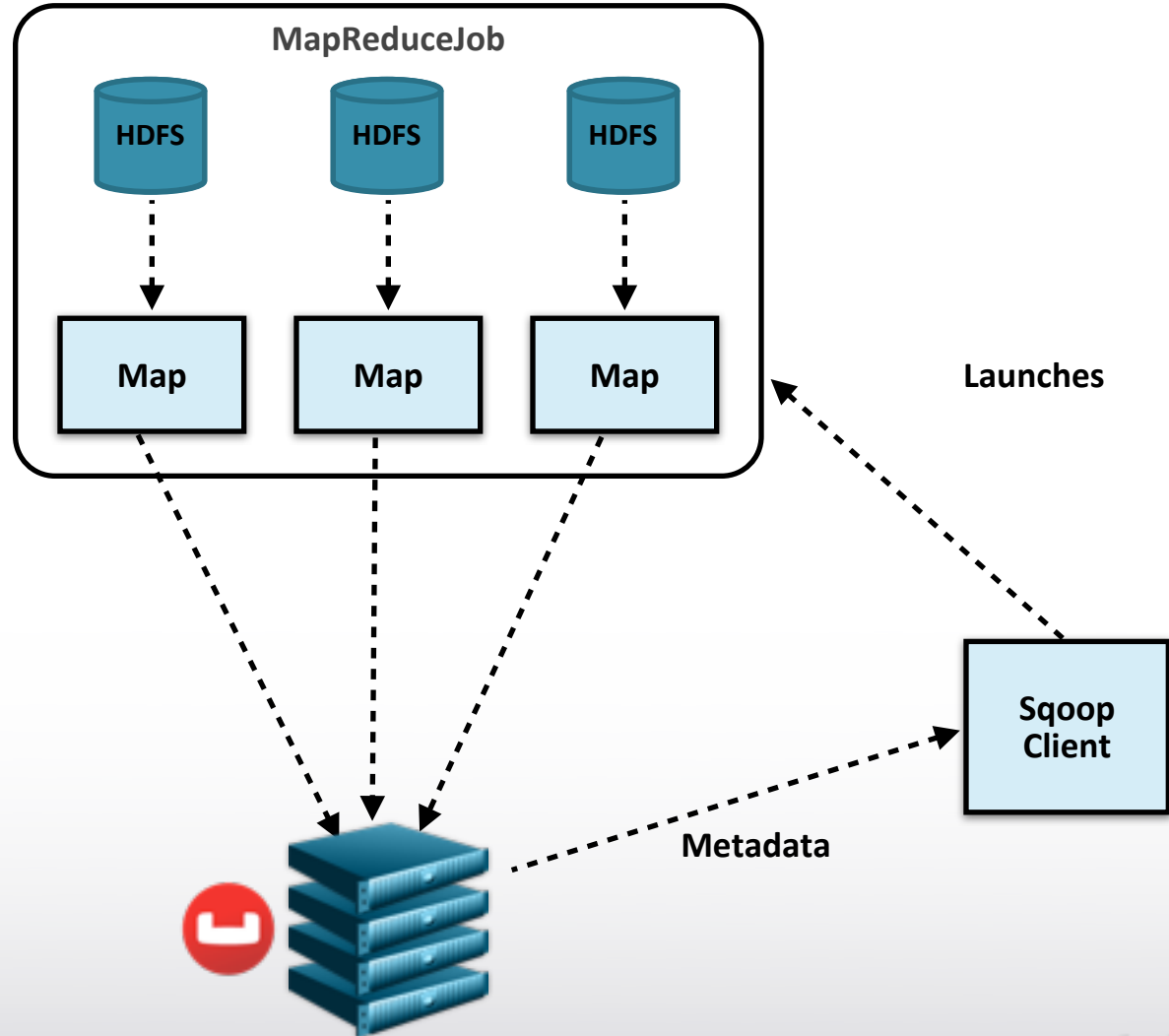


Sqoop : Export



```
sqoop export --connect http://localhost:8091/pools  
--table DUMP  
--export-dir /user/hive/profiles/recommendation  
--username social
```

Sqoop : Export





Demonstration





Couchbase

Couchbase Server Core Principles



Easy Scalability

Grow cluster without application changes, without downtime with a single click



Consistent High Performance

Consistent sub-millisecond read and write response times with consistent high throughput



Always On 24x365

No downtime for software upgrades, hardware maintenance, etc.



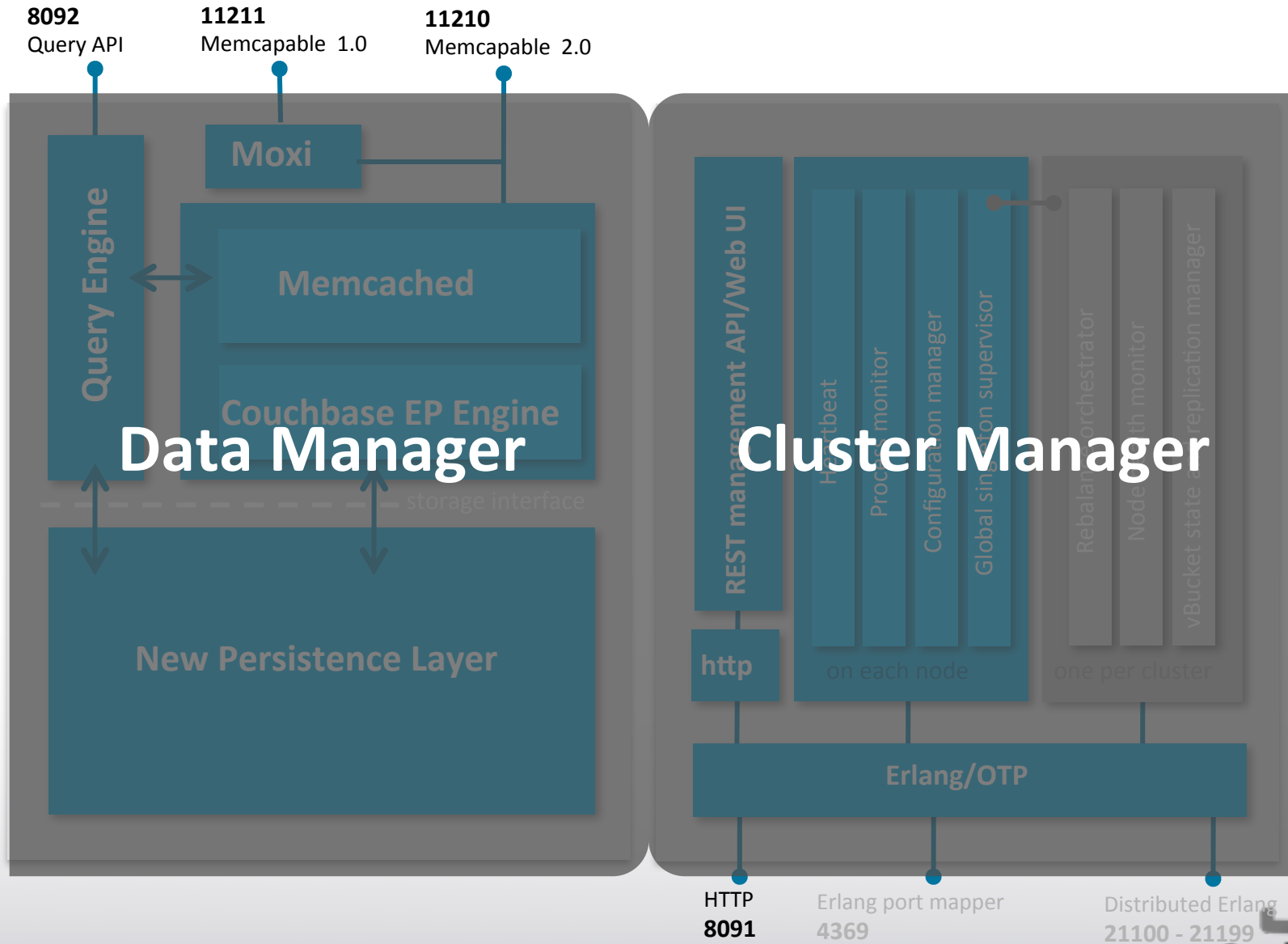
Flexible Data Model

JSON document model with no fixed schema.

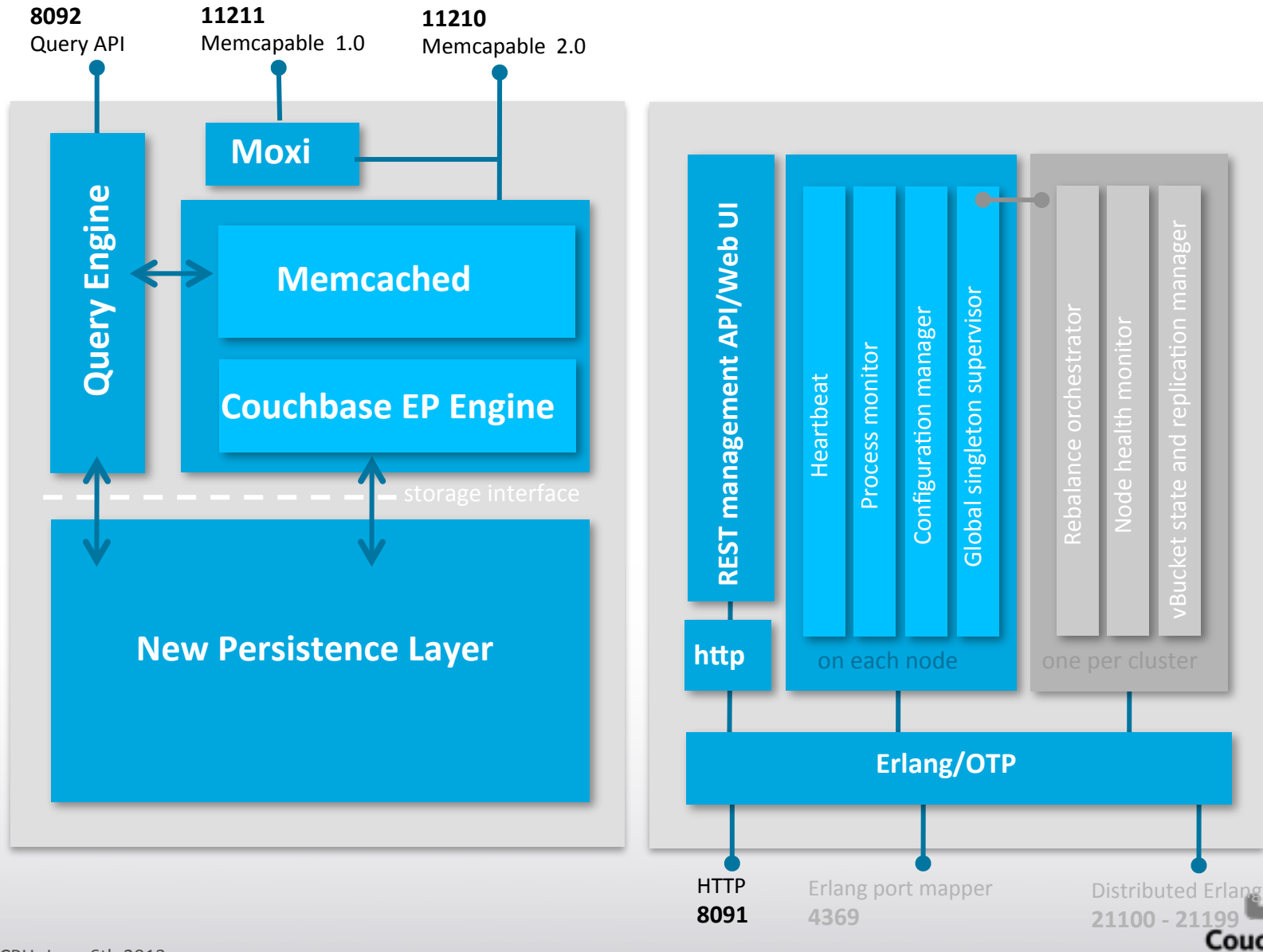
Couchbase Handles Real World Scale



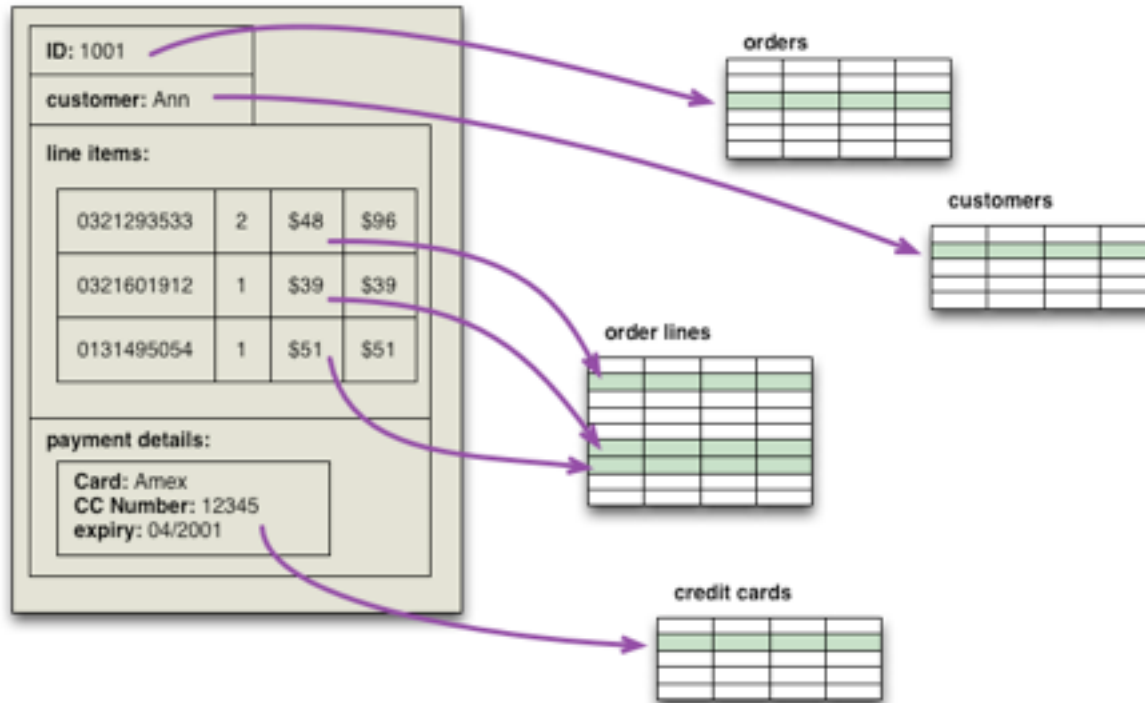
Couchbase Server 2.0



Couchbase Server 2.0



The Classic Order Entry Structure



Relational databases were not designed with clusters in mind, which is why people have cast around for an alternative. Storing aggregates as fundamental units makes a lot of sense for running on a cluster.

<http://martinfowler.com/bliki/AggregateOrientedDatabase.html>

Aggregate by Comparison

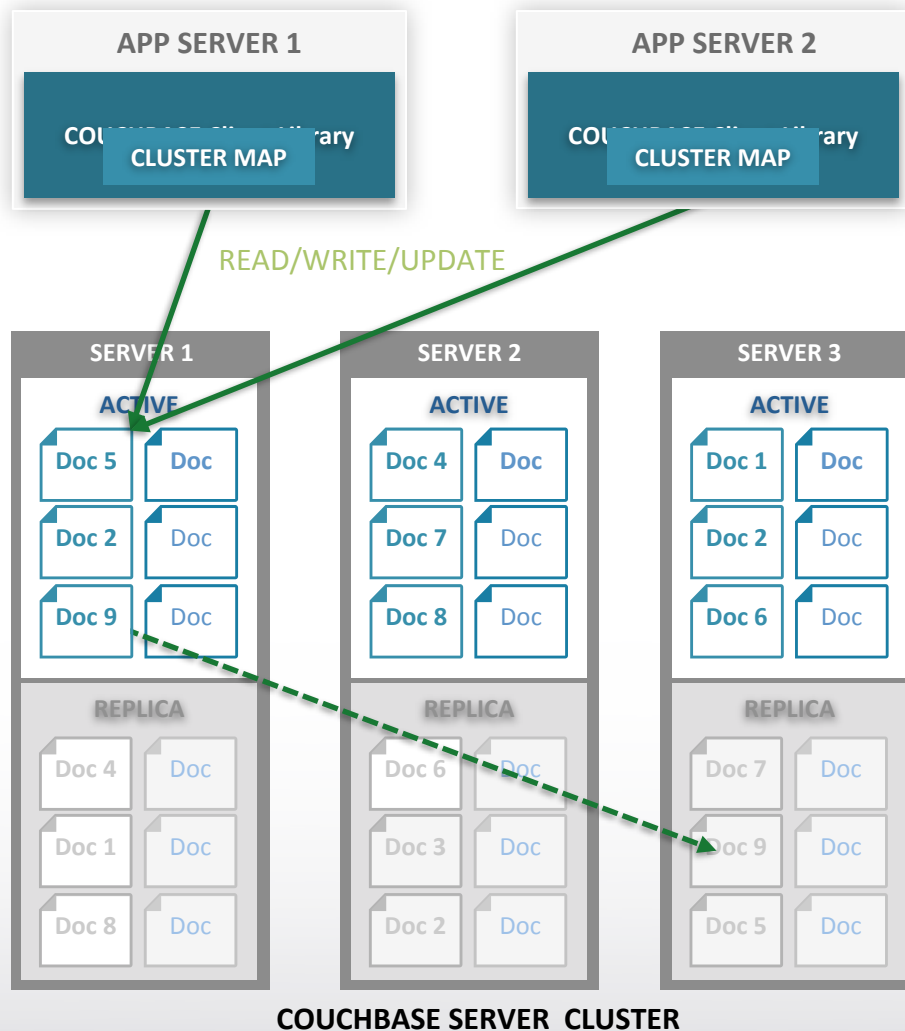
ID: 1001			
customer: Ann			
line items:			
0321293533	2	\$48	\$96
0321601912	1	\$39	\$39
0131495054	1	\$51	\$51
payment details:			
Card: Amex CC Number: 12345 expiry: 04/2001			



```
o::1001
{
  uid: "ji22jd",
  customer: "Ann",
  line_items: [
    { sku: 0321293533, quan: 3, unit_price: 48.0 },
    { sku: 0321601912, quan: 1, unit_price: 39.0 },
    { sku: 0131495054, quan: 1, unit_price: 51.0 }
  ],
  payment: {
    type: "Amex",
    expiry: "04/2001",
    last5: 12345
  }
}
```

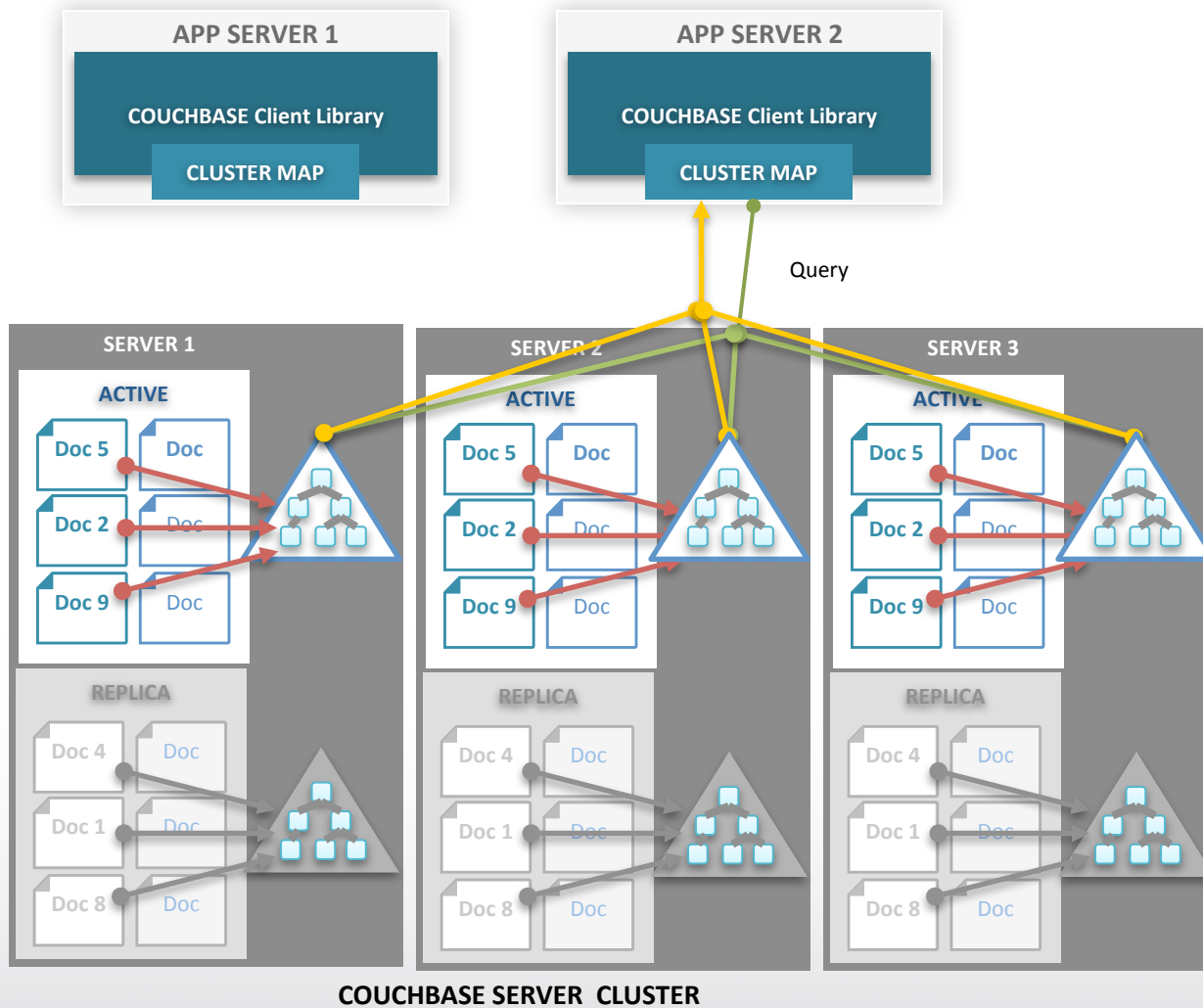
- Easy to distribute data
- Makes sense to application programmers

Basic Operations



- Docs distributed evenly across servers
- Each server stores both active and replica docs
Only one server active at a time
- Client library provides app with simple interface to database
- Cluster map provides map to which server doc is on
App never needs to know
- App reads, writes, updates docs
- Multiple app servers can access same document at same time

Indexing



- Indexing work is distributed amongst nodes
- Large data set possible
- Parallelize the effort
- Each node has index for data stored on it
- Queries combine the results from required nodes



Demonstration



Map Reduce ...



- Deal with “Big Data”
- “More” is better than “Faster”
- Batch Oriented
- Usually used to “extract/transform” data
- Fully distributed
 - Map, Shuffle, Reduce

- Distributed
 - Executed where the document is
- Deal with “indexing” data
- As fast as possible
- Use to query the data in the Database



Conclusion

- **Big Data and Big Users working together**
- **Use Hadoop to store “everything”**
 - Batch oriented
 - Complex data processing
 - MapReduce
- **Expose a subset of the dataset to your application**
 - Real time analytics
 - Low latency
 - Simple data interactions and queries

Q&A



@tgrall

tug@couchbase.com

We're Hiring! couchbase.com/careers



Q&A

