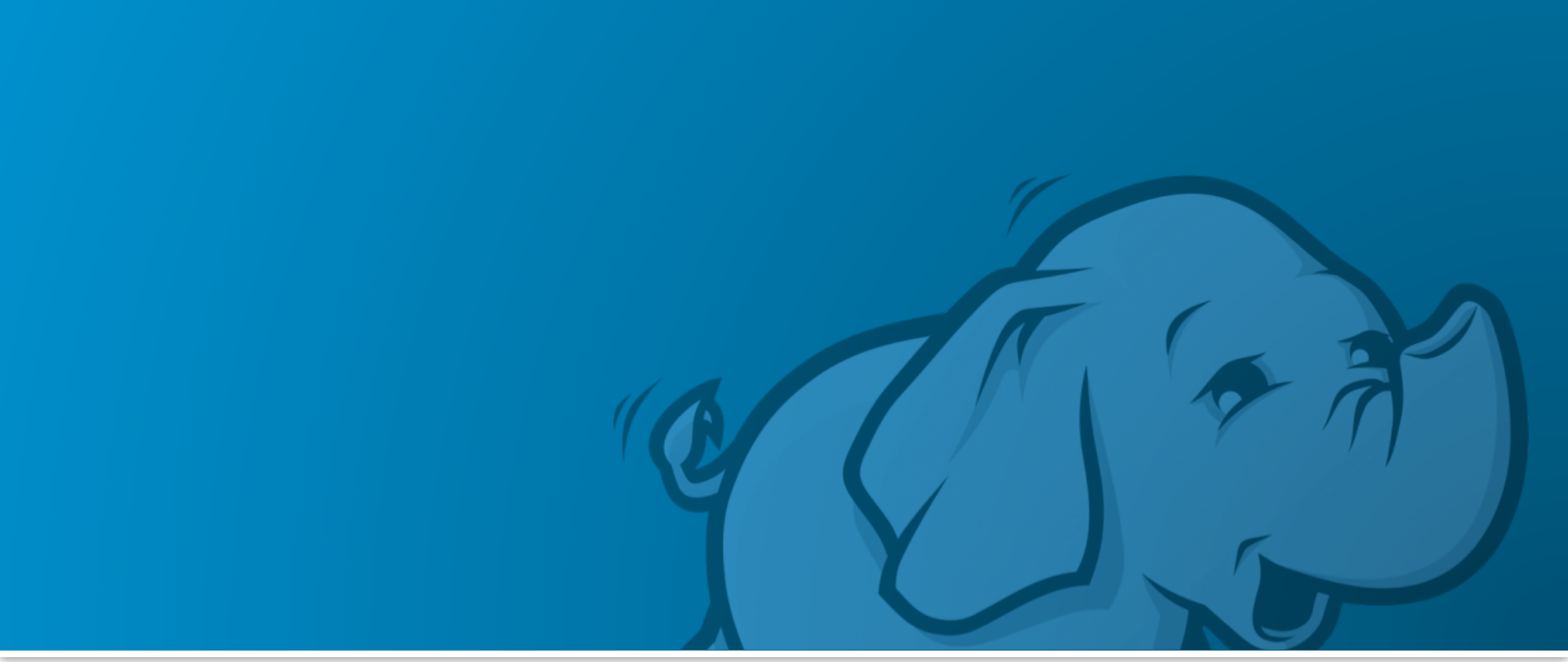


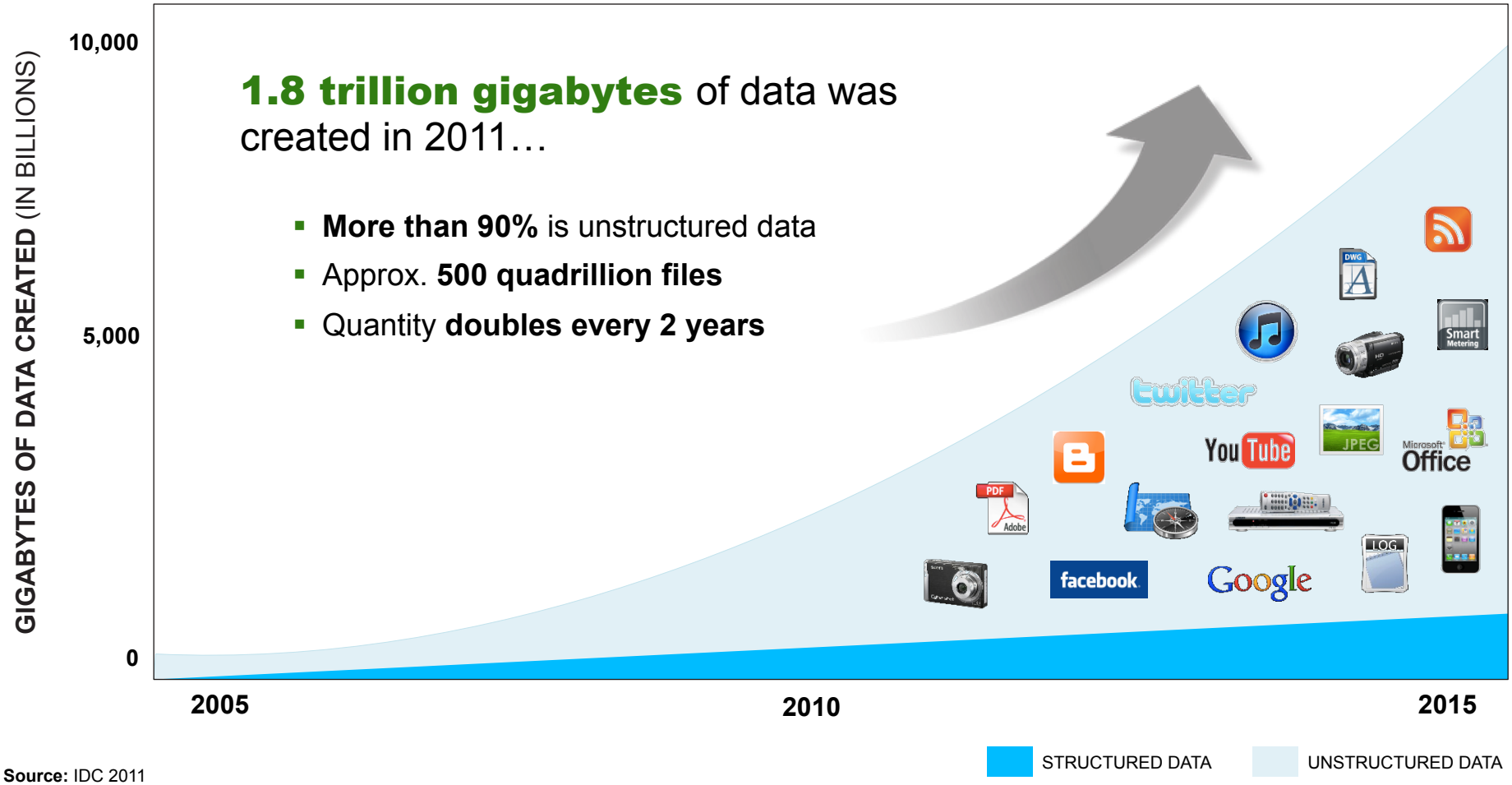
20 March 2012

What's Happening to Databases?

Mike Olson | CEO, Cloudera

cloudera

Explosive Data Growth



Source: IDC 2011

cloudera

The ‘Big Data’ Phenomenon

Big Data Drivers:

- The proliferation of data capture and creation technologies
- Increased “interconnectedness” drives consumption (creating more data)
- Inexpensive storage makes it possible to keep more, longer
- Innovative software and analysis tools turn data into information

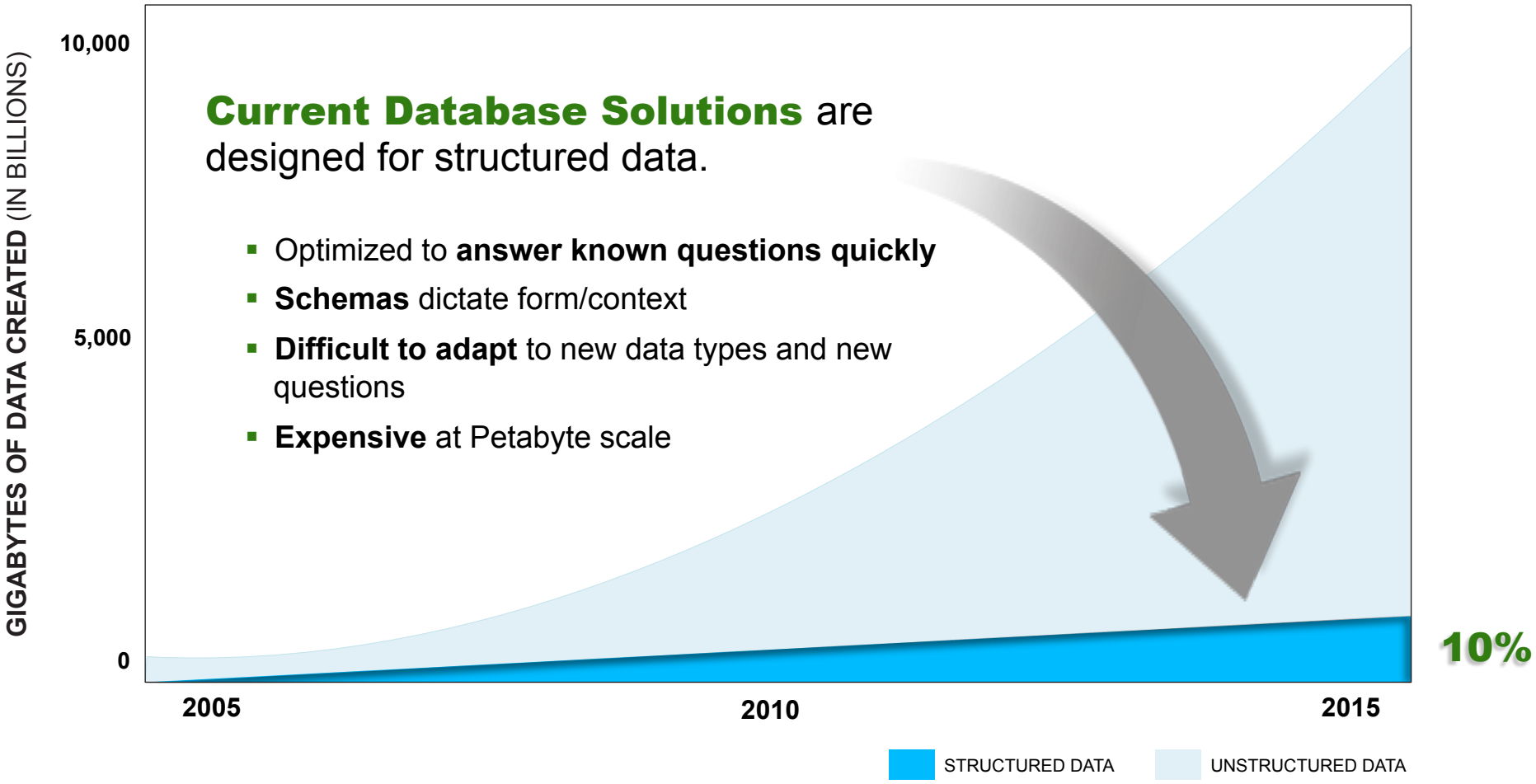
Big Data encompasses not only the **content itself**, but **how it’s consumed.**

- Every **gigabyte** of stored content can generate a **petabyte** or more of transient data*
- The information about you **is much greater** than the information you create

***Source:** IDC 2011

cloudera

The Current Solutions



cloudera

Big Data Challenges

Cost-effectively managing the **volume, velocity and variety** of data

Deriving value across
structured and unstructured data

Adapting to **context changes** and integrating
new data sources and types

cloudera

Big Data Solution Requirements

Cost-effectively manage
the volume, variety and velocity of data

Process and analyze
large, complex data sets…quickly

Flexibly adapt
to context changes and new data types

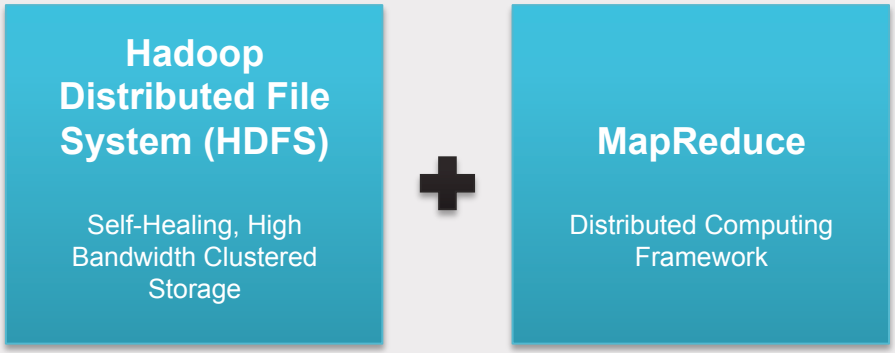
cloudera

What is Apache Hadoop?

Apache Hadoop is an open source platform for data storage and processing that is…

- ✓ Scalable
- ✓ Fault tolerant
- ✓ Distributed

CORE HADOOP SYSTEM COMPONENTS



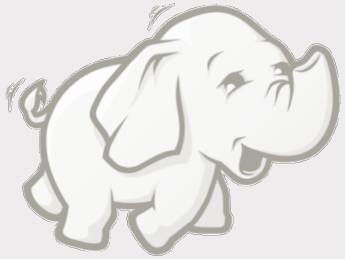
Provides storage and computation in a single, scalable system.

cloudera

What is Apache Hadoop?

Apache Hadoop is a platform for data storage and processing that is…

- ✓ Scalable
- ✓ Fault tolerant
- ✓ Open source



CORE HADOOP COMPONENTS

Hadoop Distributed File System (HDFS)

File Sharing & Data Protection Across Physical Servers

+

MapReduce

Distributed Computing Across Physical Servers

Has the Flexibility to Store and Mine Any Type of Data

- Ask questions across structured and unstructured data that were previously impossible to ask or solve
- Not bound by a single schema

Excels at Processing Complex Data

- Scale-out architecture divides workloads across multiple nodes
- Flexible file system eliminates ETL bottlenecks

Scales Economically

- Can be deployed on commodity hardware
- Open source platform guards against vendor lock

cloudera

Core Hadoop: HDFS

Self-healing, high bandwidth **clustered storage.**

HDFS breaks incoming files into blocks and stores them redundantly across the cluster.

cloudera

Core Hadoop: MapReduce

Distributed computing framework.

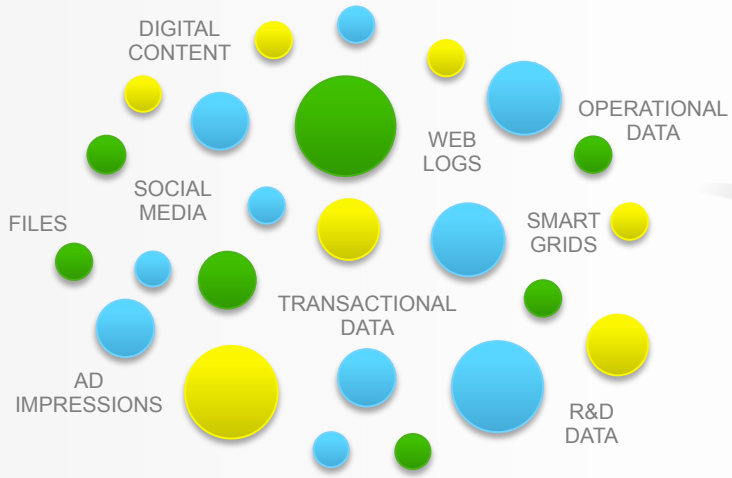
Processes large jobs in parallel across many nodes and combines the results.

cloudera

Why Was Hadoop Created?

New opportunities to derive value from all your data.

Exploding Data Volumes & Types



It's difficult to handle data this diverse, at this scale. Traditional platforms can't keep pace.

Driving The Need For A Flexible, Scalable Solution

BIG DATA

- Any Kind
- From Any Source
- Structured & Unstructured
- At Scale

HARD PROBLEMS

- Deep Analysis
- Exhaustive & Detailed
- Sophisticated Algorithms
- Generate Results Quickly

NEW OPPORTUNITIES

- Extract More Value
- From More Data
- More Cost Effectively
- With Greater Flexibility

cloudera

Core Values of Hadoop

A platform for **all your data.**

1 Scalability

- Designed to store and process data at petabyte scale
- Scale-out architecture increases capacity and processing power linearly
- Perform operations in parallel across the entire cluster

2 Flexibility

- Store data in any format – free from rigid schemas
- Define context at the time you ask the question
- Process and analyze data using virtually any programming language

3 Economics

- Build out your cluster on your hardware of choice
- Open source software guards against vendor lock-in
- Wide integration ensures investment protection

cloudera

The Value of Open Source

No Vendor Lock-In

Investments in skills, services and hardware are preserved regardless of vendor choice

Community Development

Hadoop and related projects are expanding at a rapid pace

Rich Ecosystem

Dozens of complementary hardware, software and services firms

cloudera
NETEZZA
IBM
INFORMATICA
VERTICA
hp invent
talend
open data solutions
MicroStrategy
TERADATA
Raising Intelligence

cloudera

Hadoop and Databases

Databases
"Schema-on-Write"

- Schema must be created before any data can be loaded
- An explicit load operation has to take place which transforms data to DB internal structure
- New columns must be added explicitly before new data for such columns can be loaded into the database

Hadoop
"Schema-on-Read"

- Data is simply copied to the file store, no transformation is needed
- A SerDe (Serializer/Deserlizer) is applied during read time to extract the required columns (late binding)
- New data can start flowing anytime and will appear retroactively once the SerDe is updated to parse it

1) Reads are Fast
2) Standards and Governance

1) Loads are Fast
2) Flexibility and Agility

cloudera

Hadoop and Databases

You need **both tools.**

Relational Database

Best Used For:

- Interactive OLAP Analytics (<1sec)
- Multistep ACID Transactions
- 100% SQL Compliance

Hadoop

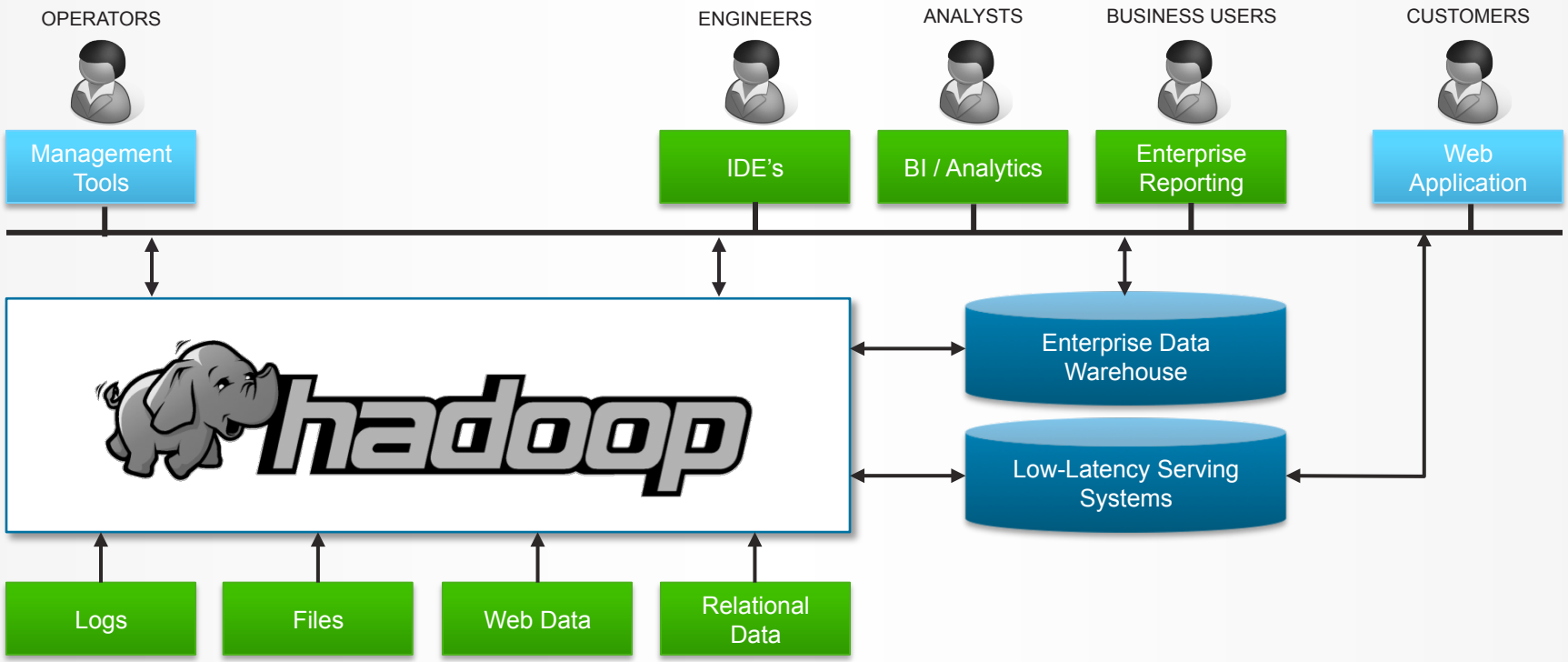
Best Used For:

- Structured or Not (Flexibility)
- Scalability of Storage/Compute
- Complex Data Processing

cloudera

Apache Hadoop in Production

How Apache Hadoop fits
into your existing infrastructure.



cloudera

The Core Values of CDH

The #1 commercial **Apache Hadoop distribution.**

1 Turn-Key System

- A complete, integrated Hadoop stack
- All component versions and dependencies are managed
- Works with a wide range of hardware, platforms and software

2 Stable and Reliable

- Thoroughly tested by the Cloudera QA team
- Fully documented and supported*
- Proven at scale in dozens of enterprise environments

3 Completely Open Source

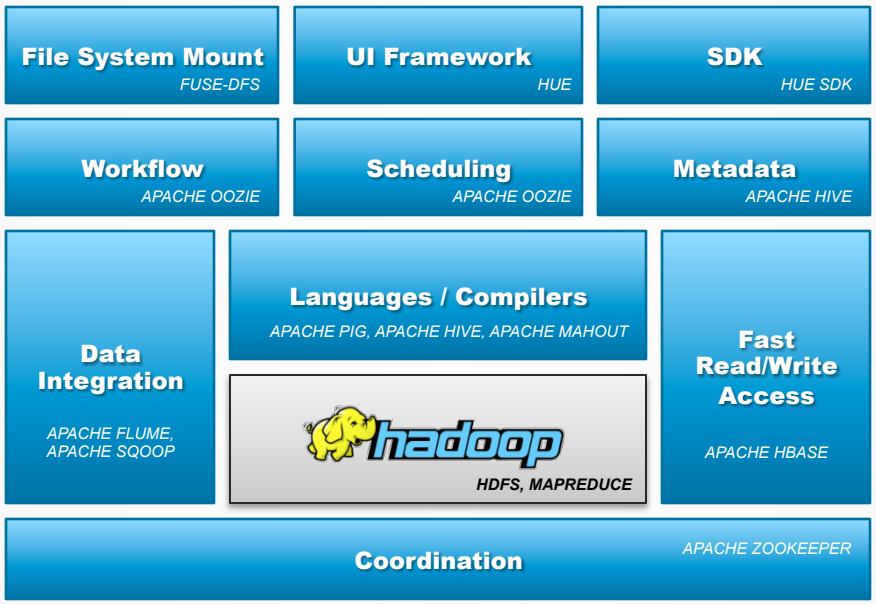
- Incorporates only mainline releases from the open source community
- No forks or proprietary underpinnings
- FREE to download

* Support is delivered through a Cloudera Enterprise subscription

cloudera

The Core Values of CDH

A **Hadoop system** with everything you need for production use.



Storage

Computation

Integration

Coordination

Access

cloudera

Hadoop Use Cases

Two **Core Use Cases**
Applied Across Verticals

1

2

ADVANCED ANALYTICS

DATA PROCESSING

INDUSTRY TERM	VERTICAL	INDUSTRY TERM
Social Network Analysis	Web	Clickstream Sessionization
Content Optimization	Media	Engagement
Network Analytics	Telco	Mediation
Loyalty & Promotions Analysis	Retail	Data Factory
Fraud Analysis	Financial	Trade Reconciliation
Entity Analysis	Federal	SIGINT
Sequencing Analysis	Bioinformatics	Genome Mapping

cloudera

WHY NOSQL?

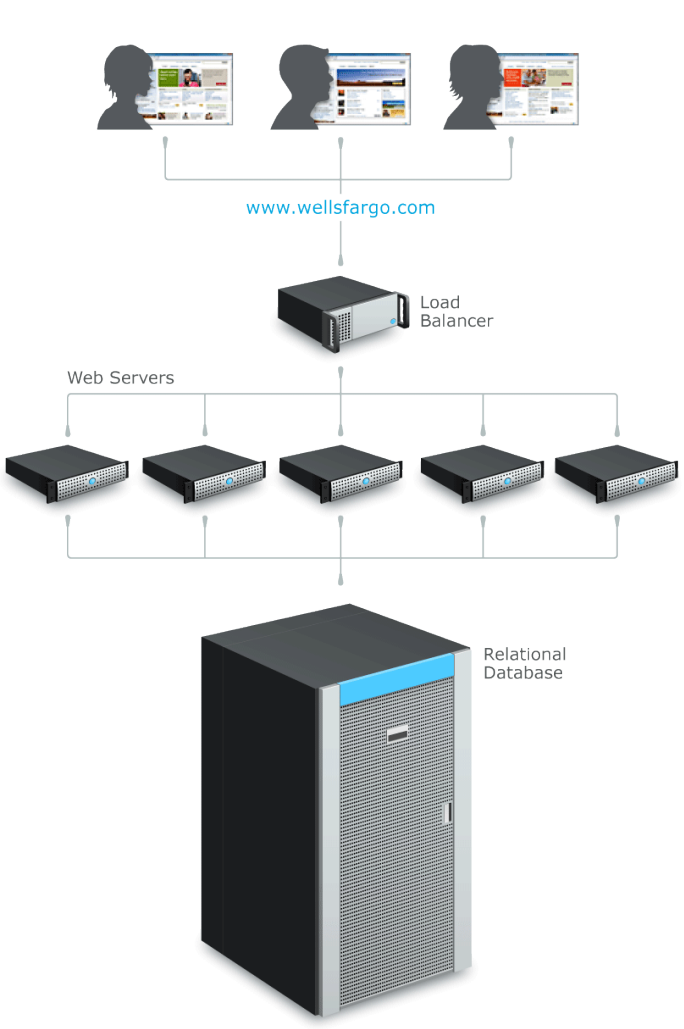
"Zynga's games serve over 235 million active users per month. We depend on technology from Couchbase to make that possible. We have *improved the performance and availability of our games while reducing hardware and administration costs.* We will continue to transition our data from relational databases to Couchbase technology."

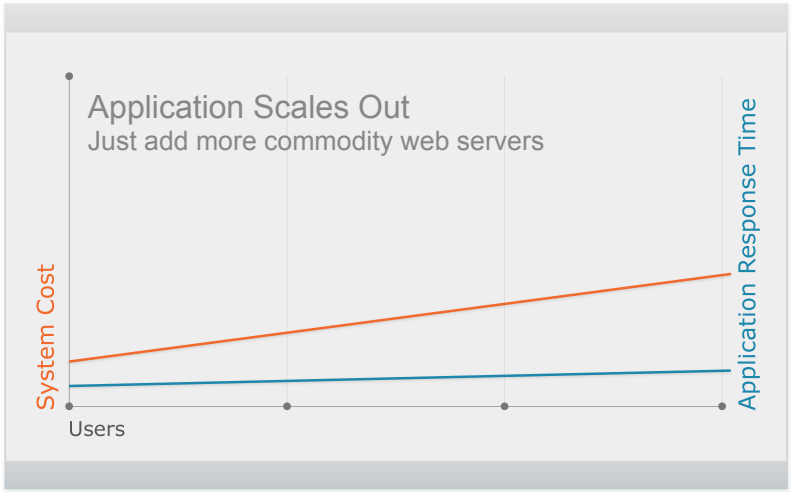
Cadir Lee
Chief Technology Officer, Zynga

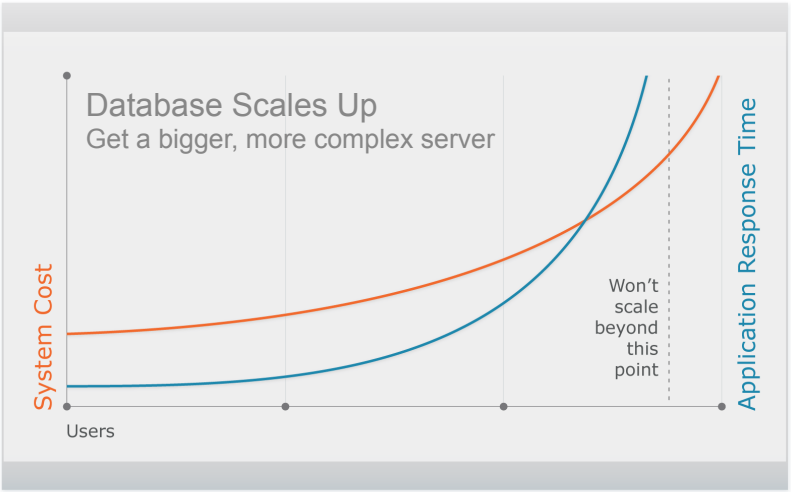
Interactive software – then and now

	Circa 1975 "Online Applications"	Circa 2011 "Interactive Web Applications"
Users	2,000 "online" users = End Point	2,000 "online" users = Starting Point
	Static user population	Dynamic user population
Applications	Business process automation	Business process innovation
	Highly structured data records	Structured, semi-structured and unstructured data
Infrastructure	Data networking in its infancy	Universal high-speed data networking
	Centralized compting (Mainframes and minicomputers	Distributed computing (Network servers and virtual machines)
	Memory scarce and expensive	Memory plentiful and cheap

Web application architecture







Lacking market solutions, users forced to invent

Bigtable
November 2006

Dynamo
October 2007

Cassandra
August 2008

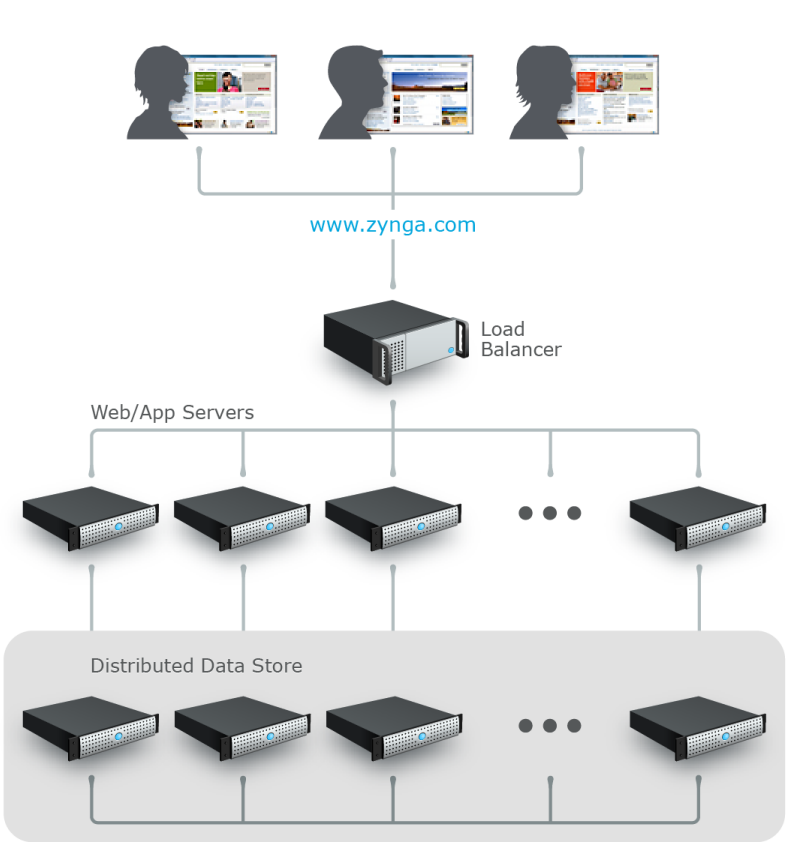
Voldemort
February 2009

Common characteristics of these "NoSQL" technologies

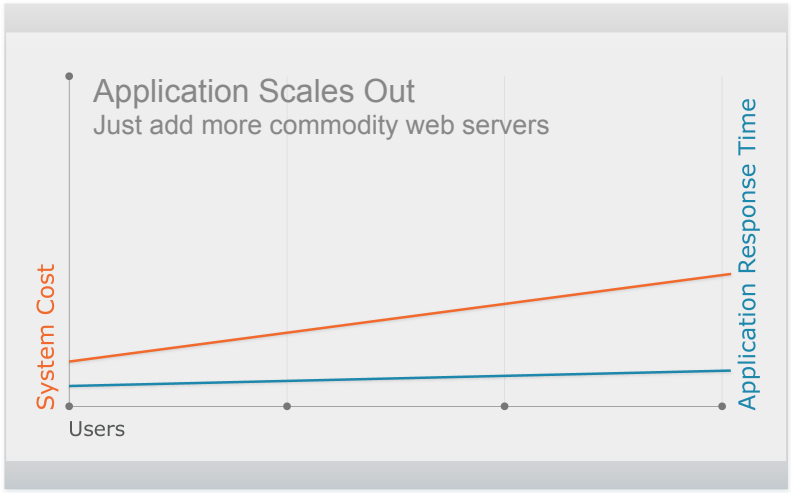
- No schema required before inserting data
- No schema change required to change data format
- Auto-sharding without application participation
- Distributed query support
- Data replication across servers and regions

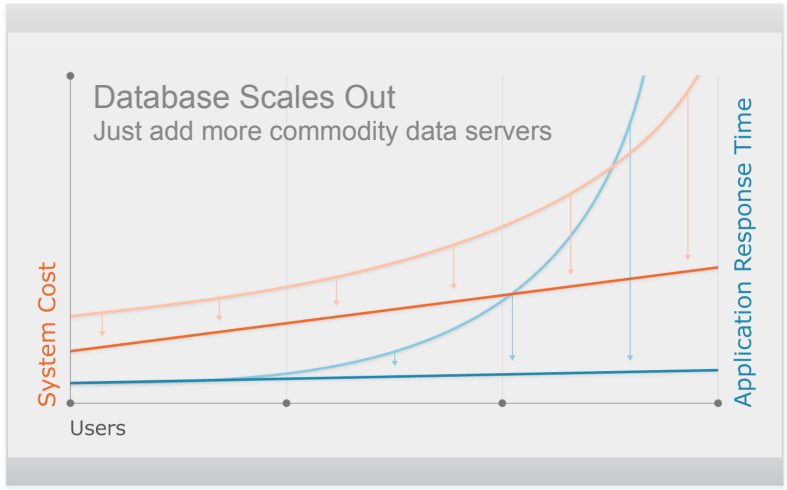
Very few organizations want to (fewer can) build and maintain database technology. Couchbase was founded to create packaged, commercially-supported NoSQL database products.

Data Layer Matches Application Logic Tier Architecture



- Horizontally scalable with auto-sharding
- High performance at web scale
- Schema-less for flexibility

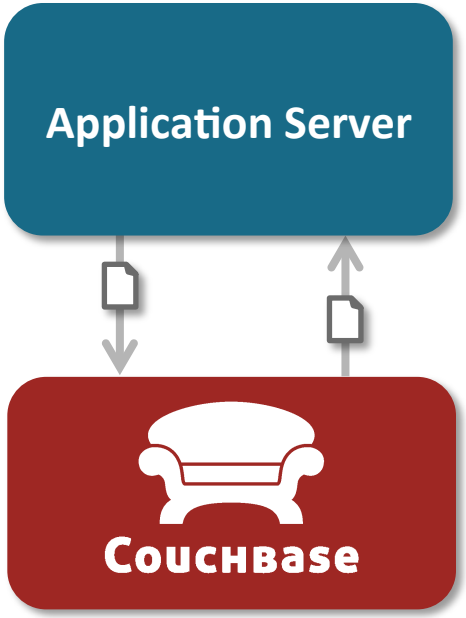




Scaling out flattens the cost *and* performance curves

Couchbase is a “document-oriented” NoSQL database

Simple. Flexible. Adjust to changing data management requirements with ease.




```
{
“UUID”: “21f7f8de-8051-5b89-86
“Time”: “2011-04-01T13:01:02.42
“Server”: “A2223E”,
“Calling Server”: “A2213W”,
“Type”: “E100”,
“Initiating User”: “dsallings@spy.net”,
“Details”:
        {
        “IP”: “10.1.1.22”,
        “API”: “InsertDVDQueueItem”,
        “Trace”: “cleansed”,
        “Tags”:
              [
              “SERVER”,
              “US-West”,
              “API”
              ]
        }
}
```

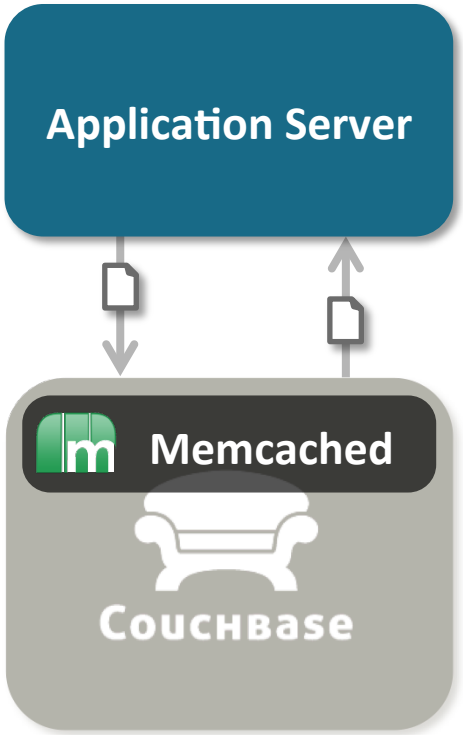

Example JSON document

Simple.

- No schema required to insert data (or change data format later). Lightweight, cross-platform document format (JSON). Efficient, native support for binary attachments.

Couchbase is *consistently* fast

Decouple application performance (user experience) from sketchy database I/O.

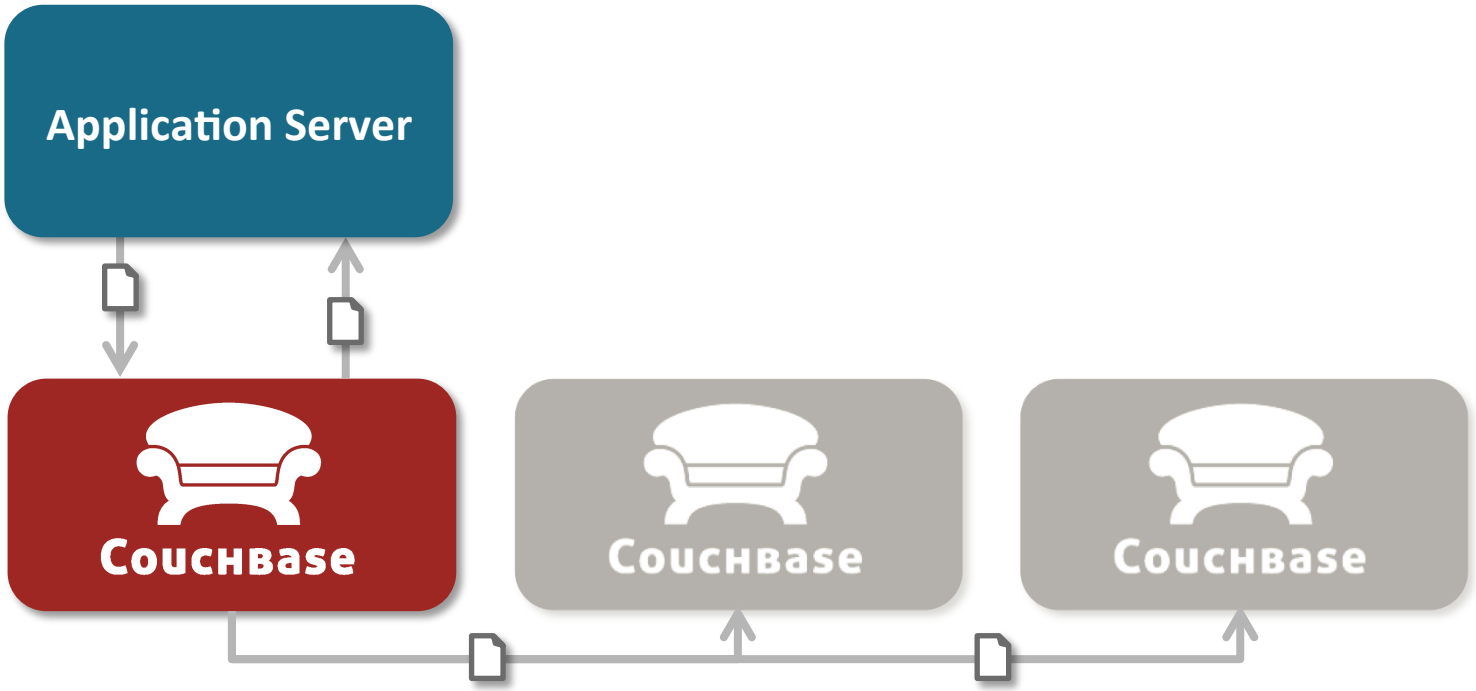


Fast.

Memcached, the most widely deployed in-memory caching technology on the planet, is built in to Couchbase enabling ***consistently*** low-latency data reads *and* writes. We wrote most of memcached.

Couchbase is elastic (scales out for increased capacity)

Grow with linear cost, constant performance and without downtime

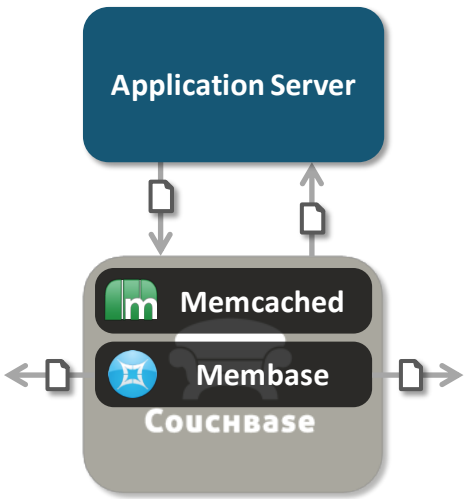


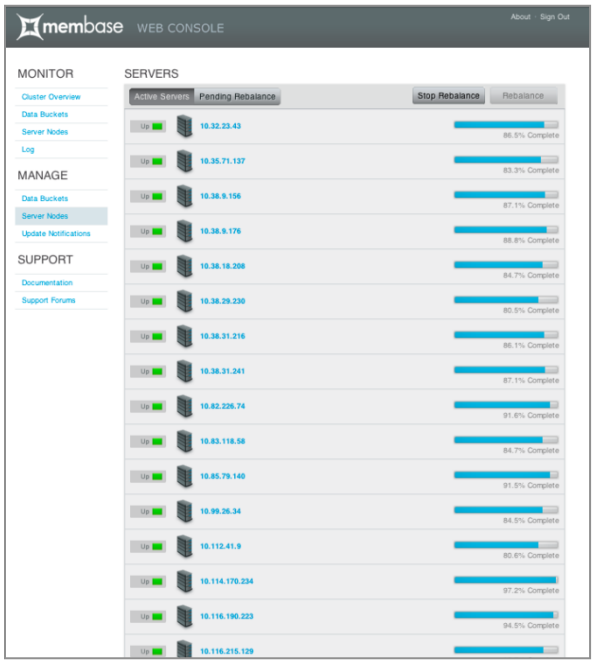
Elastic.

- Unlike other solutions, expanding (or contracting) a Couchbase cluster is effortless; and requires no application downtime.

Elasticity courtesy of Membase technology

Proven in the world's largest NoSQL production deployments





Pushbutton rebalancing of a live, 100+ node cluster

Enterprise-class cluster monitoring and administration

- Built-in Membase cluster and dataflow management technology; production proven in multi-hundred node clusters at Zynga, AOL and others. There is a BIG difference between a 20 node and a 750 node cluster; only Couchbase is as comfortable at 1 node as it is at 1000.

Watch a live, 100 node cluster rebalance at: http://blog.couchbase.com/take-a-tour-membase-server-one-dot-seven

UNQL

Say “Uncle”

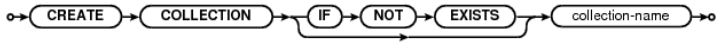
Why the need for UnQL (UnStructured Query Language)?

- NoSQL technology must mature to support more "heavy lifting" on the database itself
  - Today, application developers required to do work the database should be doing
  - Query planning, optimization and execution is not something an application developer should have to worry about
- Sufficiently expressive language required to communicate intent
- Opportunity to expand the NoSQL market, if multi-vendor support and implementation materializes
  - Education and training shared across industry
  - No vendor lock-in fears

UnQL Syntax Summary

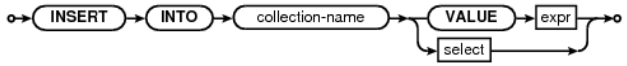
If you "KnowSQL" then you are in good shape :)

CREATE

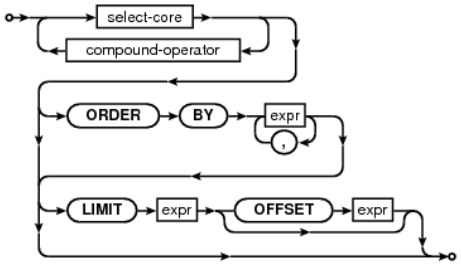


DROP

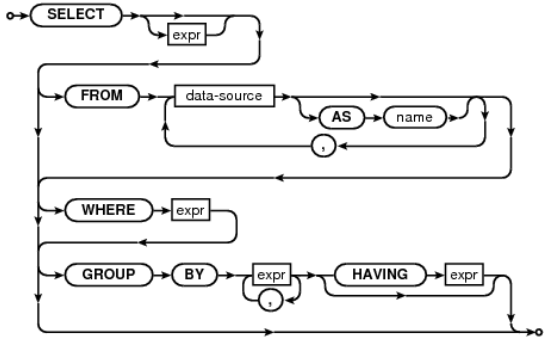
INSERT



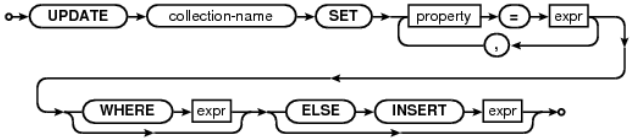
SELECT



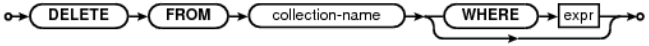
select-core



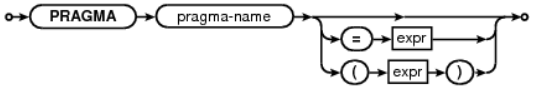
UPDATE



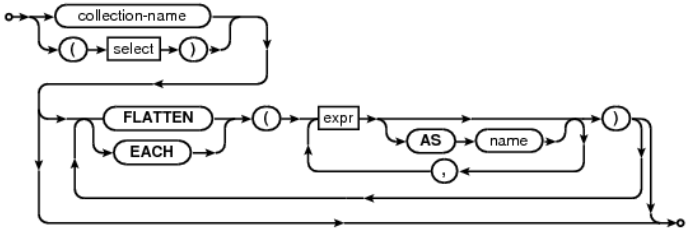
DELETE



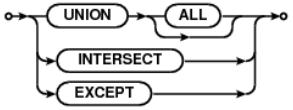
PRAGMA



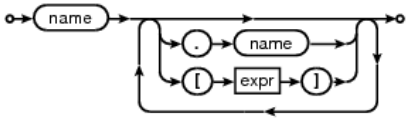
data-source



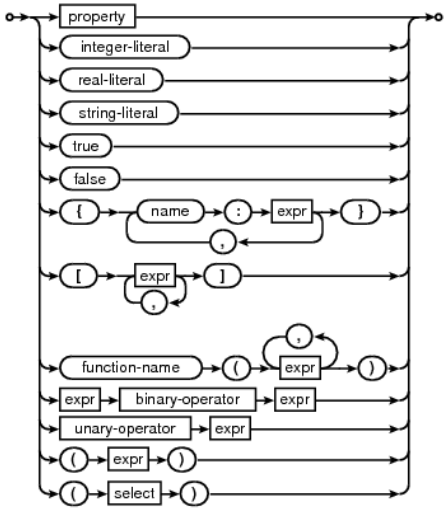
compound-operator



property



expr



What Else?

- Explosion of NoSQL alternatives
  - All the way back to Berkeley DB, with many others: MongoDB, Cassandra, Riak, …
- Columnar Stores
  - Vertica, others
- Special-purpose offerings
  - Matrix databases, graph databases, …
- Darwin at work!

cloudera

Thank You!

mike.olson@cloudera.com

@mikeolson

cloudera