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

Unstructured and Semi-Structured Data

Text, Log Files, Click Streams, Blogs, Tweets, Audio, Video, etc.

Structured Data

More Flexible Data Model Required

$30B Database Market Being Disrupted

2013: 95% Relational Technology, <5% NoSQL Technology

2027: <50% Relational Technology, >50% NoSQL Technology

All new database growth will be NoSQL
Operational vs. Analytic Databases

Real-time, Interactive Databases

- NoSQL

Fast access to data

Analytic Databases

- Hadoop

Get insights from data

Couchbase
Mongo

Cloudera
Hortonworks
Lack of flexibility/rigid schemas: 49%
Inability to scale out data: 35%
Performance challenges: 29%
Cost: 16%
All of these: 12%
Other: 11%

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

1. events
2. profiles, campaigns
3. profiles, real time campaign statistics

40 milliseconds to respond with the decision.
Moving Parts

- Couchbase Server Cluster
- sqoop import
- sqoop export
- Hadoop Cluster
- flume flow
- Ad Targeting Platform
- Logs
Content & Recommendation Targeting

1. events
2. user profiles
3. make recommendations

Content Oriented Site

Couchbase

Legacy Relational Database

Goto Night CPH, June 6th 2013
In order to keep up with changing needs on richer, more targeted content that is delivered to larger and larger audiences very quickly, data behind content driven sites is shifting to Couchbase.

Hadoop excels at complex analytics which may involve multiple steps of processing which incorporate a number of different data sources.
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

```bash
sqoop import --connect jdbc:mysql://rdbms1.demo.com/CRM
   --table customers
```
sqoop export --connect jdbc:mysql://rdbms1.demo.com/ANALYTICS
  --table sales
  --export-dir /user/hive/warehouse/zip_profits
  --input-fields-terminated-by '\0001'
Sqoop: Import

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

MapReduceJob

HDFS → Map
HDFS → Map
HDFS → Map

Metadata

Sqoop Client

Launches
Demonstration
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 Server 2.0

Data Manager
- Query Engine
- Moxi
- Memcached
- Couchbase EP Engine
- New Persistence Layer

Cluster Manager
- REST management API/Web UI
- Heartbeat
- Process monitor
- Configuration manager
- Global singleton supervisor
- Rebalance orchestrator
- Node health monitor
- vBucket state replication manager

APIs
- Query Engine: 8092
- Memcached: 11210, 11211

Storage Interface
- Erlang/OTP
- Storage interface

Ports
- HTTP 8091
- Erlang port mapper 4369
- Distributed Erlang 21100 - 21199

Monday, June 10, 13
Couchbase Server 2.0

- Couchbase EP Engine (11211 Memcached 1.0, 11210 Memcached 2.0)
- Moxi
- Couchbase Server 2.0
- Heartbeat
- Process monitor
- Configuration manager
- Global singleton supervisor
- Rebalance orchestrator
- Node health monitor
- vBucket state and replication manager
- REST management API/Web UI
- Query Engine
- New Persistence Layer
- Storage interface
- Query API (8092)
- Memcapable 1.0 (11211)
- Memcapable 2.0 (11210)
- HTTP (8091)
- Erlang port mapper (4369)
- Distributed Erlang (21100 - 21199)

Goto Night CPH, June 6th 2013

Monday, June 10, 13
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

- Easy to distribute data
- Makes sense to application programmers

```
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
  }
}
```
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
We’re Hiring!  couchbase.com/careers

Q&A

@tgrall
tug@couchbase.com