

# NoSQL databases

Hbase & friends

Social Geek #3

# Why NoSQL ?

"Relational databases give you too much. They force you to twist your object data to fit a RDBMS..." Jon Travis

"Web 2.0 companies can take chances and they need scalability," Johan Oskarsson

"When you have these two things in combination, it makes [NoSQL] very compelling."

# What is NoSQL?

1. Not a database!

- Dynamo: "highly available key-value store."
- BigTable: "distributed storage system for managing structured data."

2. They can store enormous amounts of data

# Major open source players

## Hbase

- <http://hadoop.apache.org/hbase/>

## Cassandra

- <http://incubator.apache.org/cassandra/>

## CouchDB:

- <http://couchdb.apache.org/>

## MongoDB:

- <http://www.mongodb.org>

A complete list at:

<http://www.metabrew.com/article/anti-rdbms-a-list-of-distributed-key-value-stores/>

# What is Hbase?

HBase is a...

- Sorted
- Distributed,
- Column-Oriented,
- Multi-Dimensional,
- Highly-Available,
- High-Performance,
- Persisted Storage System



# What HBase is not...

## A SQL Database!

- Not relational
- No joins
- No sophisticated query engine
- No column typing
- No SQL, no ODBC/JDBC, Crystal Reports, etc.
- No transactions
- No secondary indices

# How to choose?

- 99% of the applications never need this level of scalability
- premature optimization is the root of all evil
- be pragmatic

Case study

Feed Aggregator



# Hbase schema

`Feeds` table - columns `Content:` and `Meta:`  
raw content in `Content:raw`

`Urls` table - columns `Content:` and `Meta:`  
feed content in `Content:raw`

`UrlsIndex` table - columns `Url:`  
keyformat: <cat>/<timestamp\_w3c\_format>[+00x]

# Source code

Open source:

<http://github.com/andreisavu/feedagggregator>

Feel free to fork and play with the source code.

Tags: Python, Thrift, Hbase 0.20, feedparser

# Resources #1 - Hbase & Bigtable

Hbase & Bigtable:

<http://hadoop.apache.org/hbase/>

<http://hadoop.apache.org/>

<http://labs.google.com/papers/bigtable.html>

<http://labs.google.com/papers/mapreduce.html>

# Resources #2 - Python & Thrift

<http://www.python.org/>

<http://www.diveintopython.org/>

<http://incubator.apache.org/thrift/>

# Resources #3 - Samples

Yet another feed aggregator:

<http://github.com/andreisavu/feedaggregator>



# Questions ?

Email: [contact@andreisavu.ro](mailto:contact@andreisavu.ro)

Twitter: <http://twitter.com/andreisavu>

Web: <http://www.andreisavu.ro>