NoSQL Use Cases

Jeremy Mikola

@jmikola

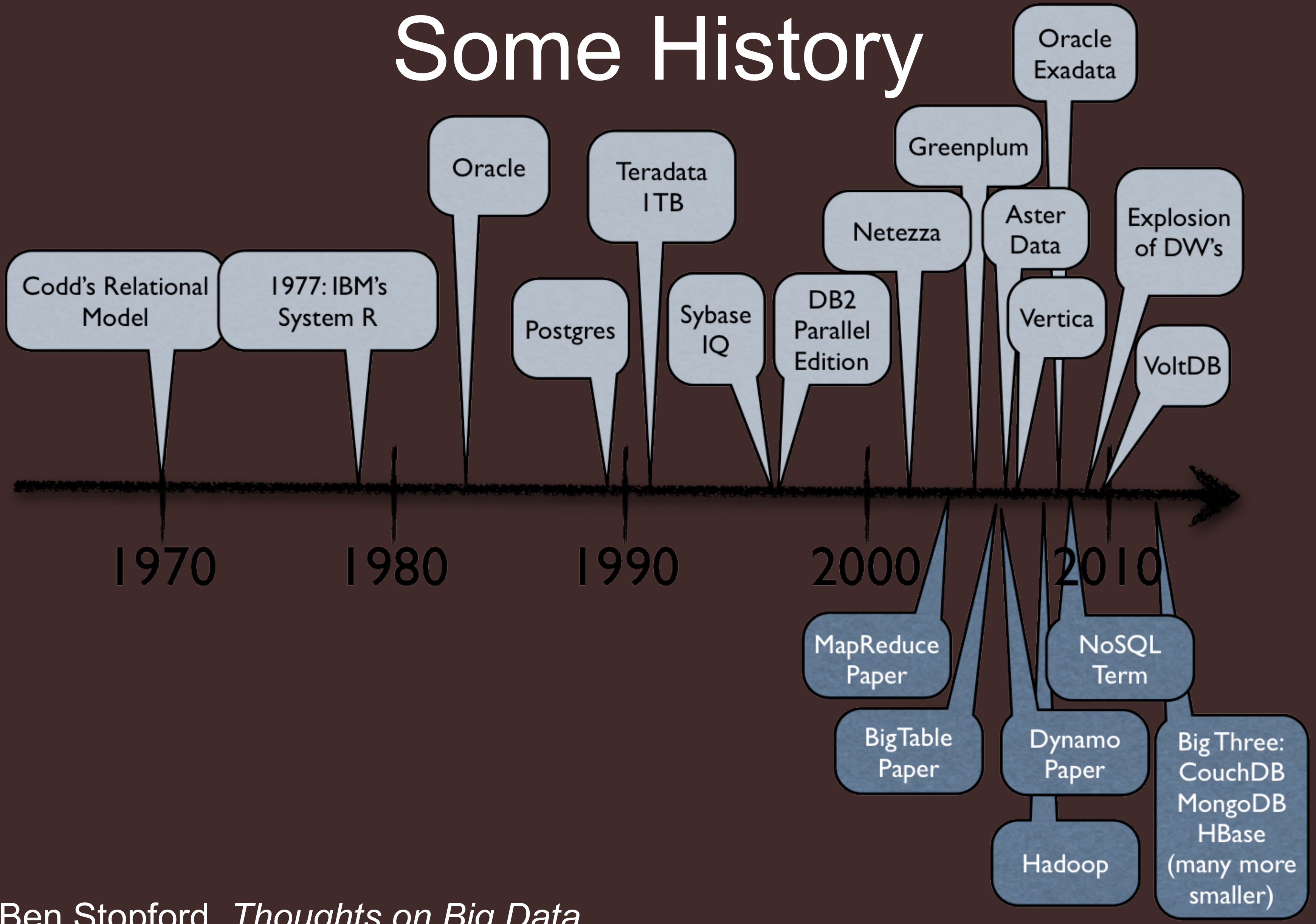
10gen | the MongoDB company

mongoDB

10gen | the MongoDB company

- Develops MongoDB and its drivers as OSS
- Professional support, training, and consulting
- Host and sponsor of conferences, user groups
- Offices: NYC, Palo Alto, London, Dublin, Sidney

Some History

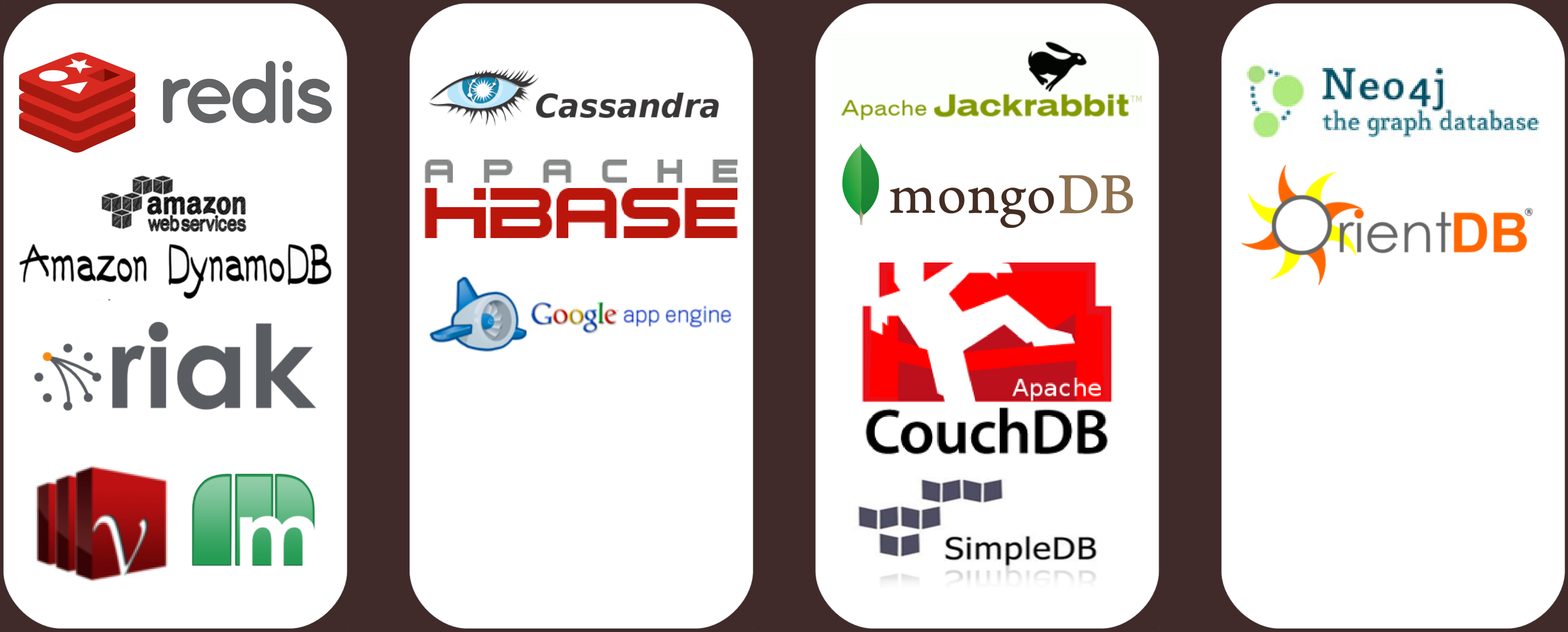


Ben Stopford, *Thoughts on Big Data*
http://www.benstopford.com/2012/06/30/thoughts-on-big-data-technologies-part-1/

10gen | the MongoDB company

mongoDB

What is NoSQL?



Key/Value

BigTable

Document

Graph

Key/Value Stores

- Maps arbitrary keys to values
- No knowledge of the value's format
- Completely schema-less
- Implementations
  - Eventually consistent, hierarchal, ordered, in-RAM
- Operations
  - Get, set and delete values by key

BigTable

- Sparse, distributed data storage
- Multi-dimensional, sorted map
- Indexed by row/column keys and timestamp
- Data processing
  - MapReduce
  - Bloom filters

Graph Stores

- Nodes are connected by edges
- Index-free adjacency
- Annotate nodes and edges with properties
- Operations
    - Create nodes and edges, assign properties
    - Lookup nodes and edges by indexable properties
    - Query by algorithmic graph traversals

Document Stores

- Documents have a unique ID and some fields
- Organized by collections, tags, metadata, etc.
- Formats such as XML, JSON, BSON
- Structure may vary by document (schema-less)
- Operations
    - Query by namespace, ID or field values
    - Insert new documents or update existing fields

What's the common thread?

What's the common thread?

All address some limitation(s) of relational DBs

Horizontal scalability, read/write performance, schema limitations, unconventional query patterns, parallel data processing, administration, etc.

What are we looking for?

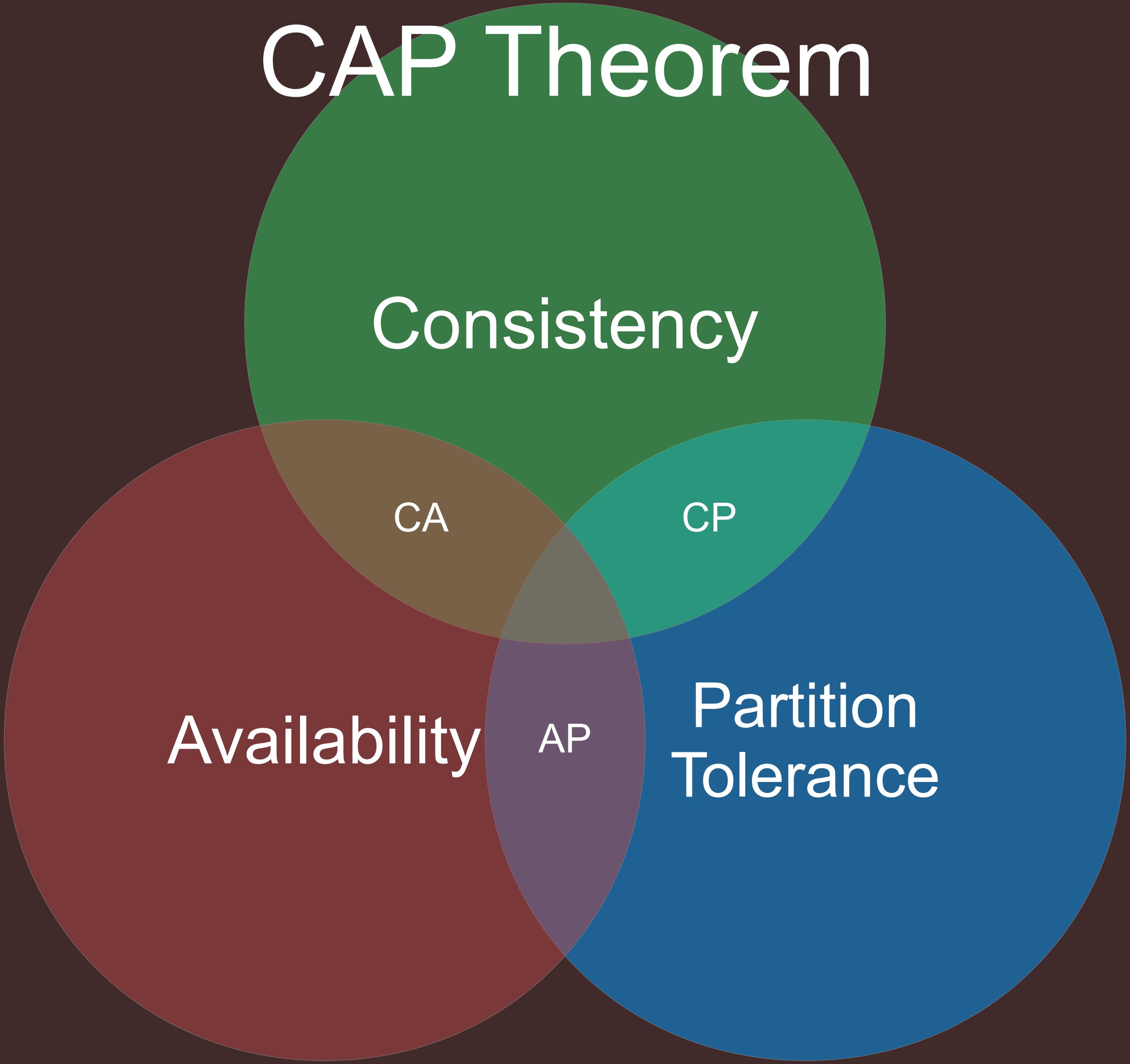
- Read/write availability and/or performance
- Avoiding a single point of failure
- Flexible schema and data types
- Ease of maintenance, administration
- Parallel computing (e.g. MapReduce)
- Supporting large data sets with room to grow
- Tunable for deployment size or functionality

Some specific needs

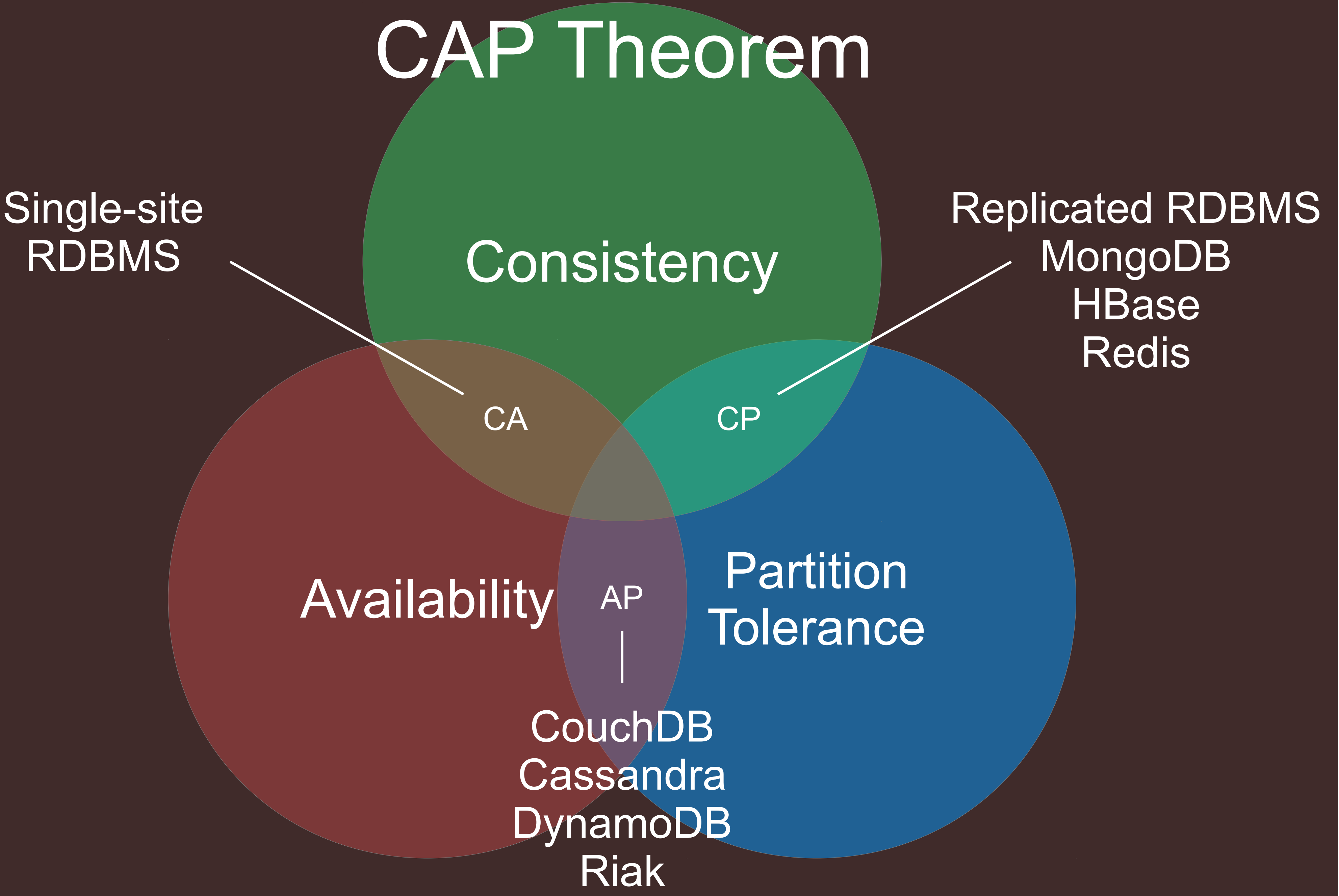
- Storing large streams of non-transactional data
  - e.g. log aggregation, ad impressions, web stats
- Syncing on/offline data (CouchBase Mobile)
- Caching results from slower data stores
  - Provide faster in-app response times
  - Denormalize results of expensive join queries
- Real-time systems (games, financial data)

What are the challenges and trade-offs?

CAP Theorem



CAP Theorem



10gen | the MongoDB company

mongoDB

> In partitioned databases, trading some consistency for availability can lead to dramatic improvements in scalability.

Dan Pritchett, *BASE: An ACID Alternative*
http://queue.acm.org/detail.cfm?id=1394128

MongoDB Philosophy

- Document data models are good
- Non-relational model allows horizontal scaling
- Provide functionality whenever possible
- Strongly consistent, durable (data is important!)
- Minimize the learning curve
    - Easy to setup and deploy anywhere
    - JavaScript and JSON are ubiquitous
    - Automate sharding and replication

Case Study: Craigslist

- 1.5 million new classified ads posted per day
- MySQL clusters
  - 100 million posts in live database
  - 2 billion posts in archive database
- Schema changes
  - Migrating the archive DB could take months
  - Meanwhile, live DB fills with archive-ready data

Case Study: Craigslist

- Utilize MongoDB for archive storage
- Average document size is 2KB
- Designed for 5 billion posts (10TB of data)
- High scalability and availability
    - New shards added without downtime
    - Automatic failover with replica sets

> We can put data into MongoDB faster than we can get it out of MySQL during the migration.

Jeremy Zawodny, software engineer at Craigslist and author of *High Performance MySQL*

http://blog.mongodb.org/post/5545198613/mongodb-live-at-craigslist

Case Study: Shutterfly

- 20TB of photo metadata in Oracle
- Complex legacy infrastructure
  - Vertically partitioned data by function
  - Home-grown key/value store
- High licensing and hardware costs

Case Study: Shutterfly

- MongoDB offered a more natural data model
- Performance improvement of 900%
- Replica sets met demand for high uptime
- Costs cut by 500% (commodity hardware)

Case Study: OpenSky

- E-commerce app built atop Magento platform
- Multiple verticals (clothing, food, home, etc.)
- MySQL data model was highly normalized
- Product attributes were not performant

Case Study: OpenSky

- Integrated MongoDB alongside MySQL
- Documents greatly simplified data modeling
  - Product attributes
  - Configurable products, bundles
  - Customer address book
- Purchases utilized MySQL transactions
- Denormalized order history kept in MySQL

Case Study: Gauges

- SaaS for real-time web analytics
- Recording time-series data in documents
  - Aggregate and display by hour, day, month, year
  - Visits, screen size, geo location, search terms, etc.
- MongoDB replica set for scalable storage
- Kestrel distributed, message queue in front
  - Highest availability, never misses a write operation

Additional Case Studies

10gen.com/customers

How Twitter Uses NoSQL

- Facebook's Scribe for log aggregation
- Hadoop for clustered data storage
  - Yahoo's Pig scripting language for querying
- Hbase for low-latency people searches
- FlockDB for social graph queries
  - Real-time, distributed, built upon MySQL
- Cassandra for data-mining and analytics

http://readwrite.com/2011/01/02/how-twitter-uses-nosql

Using the Right Tool for the Job

Questions?

10gen | the MongoDB company

mongoDB