QuABase: A Dynamic Software Engineering Knowledgebase for Building Big Data Systems

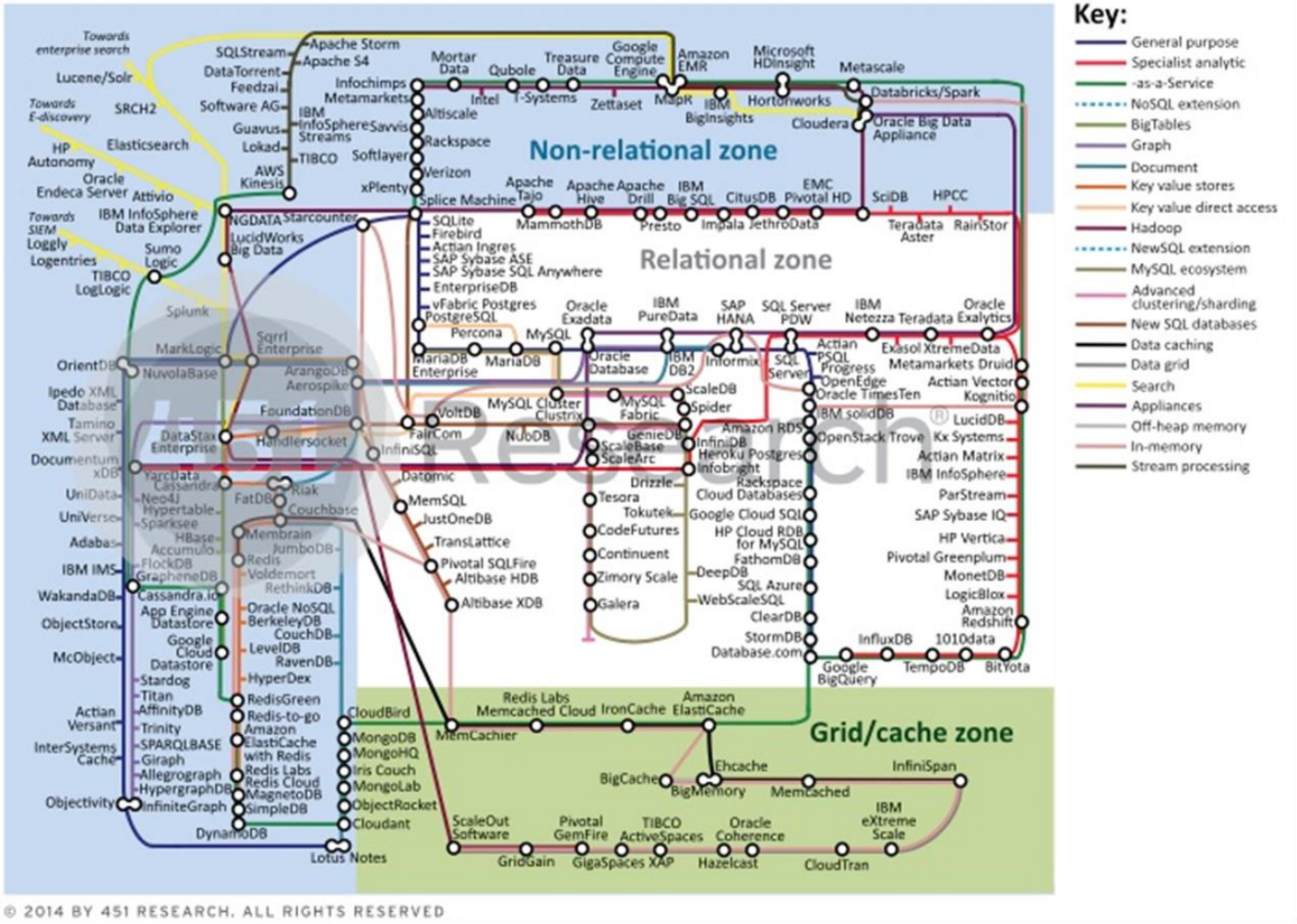
Carnegie Mellon University
Software Engineering Institute
Pittsburgh, PA 15213

Ian Gorton

Software Engineering Institute | Carnegie Mellon

Scale Changes Everything

Big Data Computational Infrastructure



http://blogs.the451group.com/information_management/2014/03/18/updated-data-platforms-landscape-map-february-2014/

Our Approach: QuABase

Semantics-based Knowledge Model

- Generic for model software architecture knowledge
- Populated with specific big data architecture knowledge

Structured knowledge capture to populate the model

Dynamic, generated and queryable content

Visualization

QuABase

Edit Data Replication: Riak Data Replication Features

Special:RunQuery/Consistency Query > Images/4/48/BigData.png > Main Page > Riak > Riak Data Replication Features

Replication Options | Failover Options

Replication Architecture: peer-to-peer

Replication for Backup: not supported

Replication across Data

Replicas Writes:

Replica Reads:

Read Repair:

Notes:

Riak uses consistent h

Riak Data Replication Features

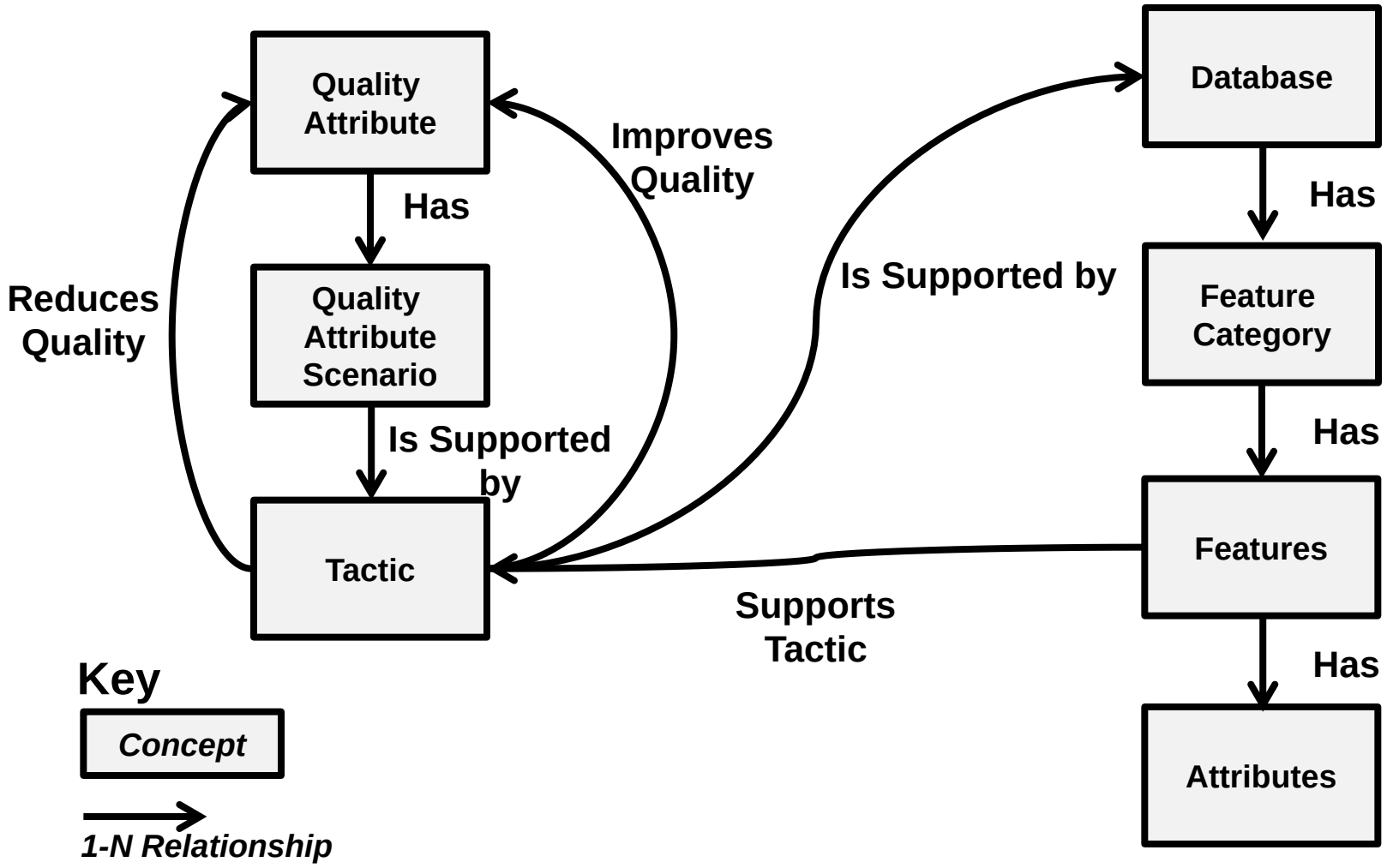
Ensure read/write quorums > Images/4/48/BigData.png > Main Page > Riak > Riak Data Replication Features

Replication Features

This section describes the range of options for configuring data replication in Riak. Replication is necessary to achieve high levels of availability in big data systems, as well as enhancing performance and scalability.

1. **Replication Architecture:** There are two basic approaches to data replication. Master-slave maintains a master copy of each data object and replicates this to 1..N other nodes. Updates typically are made to the master, although some databases allow slave updates which are coordinated transactionally with the master. With peer replication, an algorithm, typically consistent hashing, distributes N copies of each data object across different nodes. Updates may take place at any copy, and other replicas are updated either immediately or eventually depending on database configuration. In Riak, the replication architecture is peer-to-peer.
2. **Replication for Backup:** Some databases can replicate data for backup purposes. This is referred to as a *rolling backup*, and can be useful for recovering from some failure scenarios. No client traffic is sent to backup replicas, and the delay with which replication occurs can typically be configured to trade-off performance and data currency to suit application needs. In Riak, backup replicas are not suppported.
3. **Replication across Data centers:** Wide area replication across geographically distributed data centers introduces higher availability guarantees at the cost of additional resources and overheads. In Riak, replication across data centers is supported in enterprise version only.
4. **Replica Writes:** Replicated databases typically offer configuration options that enable an application to specify the number of replicas to write to, and in some cases which replicas to write to. In Riak, the following options are available: to any replica, to multiple replicas.
5. **Replica Reads:** Replicated databases typically offer configuration options that enable an application to specify the number of replicas to read from, and in some cases which replicas to read. In Riak, the following options are available: from any replica, from multiple replicas.
6. **Read Repair:** When data object replicas become inconsistent, read repair is a mechanism to make them consistent by overwriting the replica with an older values with the latest value. This feature is typically found when peer replication is used, and in Riak the options are: per query, background.

QuABase Semantic Model



Next Steps

Perform experiments/tests to validate utility of QuABase

Limited QuABase release for community testing

Extend content for several new big data technologies

- Working with CMU grad students

New project – use machine learning to automatically populate the QuABase

- With Professor Yiming Yang and students @ CMU LTI

WE'D LOVE TO COLLABORATE

More Information

http://blog.sei.cmu.edu/

OCT 21 2013 — Addressing the Software Engineering Challenges of Big Data

JAN 13 2014 — The Importance of Software Architecture in Big Data Systems

JUL 14 2014 — Four Principles of Engineering Scalable, Big Data Software Systems

AUG 11 2014 — Principles of Big Data Systems: You Can't Manage What You Don't Monitor

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6774768