MySQL Cluster Powers Leading Document Management Web Service

Document Management Web Service

Platform: Amazon EC2 and S3 Cloud Platform[1]

Database: MySQL Cluster 7.0 and 7.1

"With several years of experience of using MySQL Cluster, Docudesk has found MySQL to handily exceed our requirements for low latency, high throughput perform-ance with continuous availability, in a single solution that minimizes complexity and overall cost."

Casey Brown
DBA and Development Manager
Docudesk

DocQ
https://docq.com

Docudesk Overview

Founded in 2001 and based out of Dallas, Texas, Docudesk is a leading devel-oper of PDF document software, created with the goal of providing quality PDF software that delivers innovative features at an affordable price.

After seeing thousands of customer requests come through Docudesk over the years, it became clear that the main reason customers were creating PDF's was to create a professional looking and secure document, which they ultimately needed to share with their customers. But customers would spend time crafting the PDF documents, only to then print and fax them or send via a courier ser-vice, thereby completely eliminating the utility of the PDF in the first place and unnecessarily extending their sales cycles, contract approvals, loan agreements, and insurance claims.

To address these issues, Docudesk developed the DocQ web service to elimi-nate the limitations of sharing physical documents. DocQ is a complete paper-less business solution; providing a single place where customers can manage, archive, and send their important documents. By design, DocQ is built to sup-port not only secure business transactions but to provide the supporting facili-ties needed to store, edit, collaborate, and publish business documents. Specifically, the DocQ service is designed to deliver:

- A hassle-free way for customers to get legally signed and binding documents without printing, faxing, and overnight shipping.

- Facilitate the transition from a physical to a digital document by providing "onramp" technology enabling scanning and faxing use cases to easily work with DocQ.

- Provide a rich internet application that would give DocQ customers a solu-tion requiring no software or hardware configuration and that is much more accessible than a fax machine.

- Preserve the intent of the PDF standard by honoring document quality, porta-bility and security.

- Provide an unsurpassed level of trust and simplicity.

- Provide complementary tools and features that enable businesses to use DocQ for all paperless efforts including; mobile access, document archiving, form response, scanner integration, and document publishing.

- Design a system that meets or exceeds relevant standards; including e-Sign act of 1999 and UETA .

[1] Note: Oracle does not formally support MySQL Cluster on cloud deployment platforms.

The Business Requirements

As a totally new service, the systems architects at Docudesk had a great deal of flexibility in selecting the technology that would power the DocQ application. This flexibility extended to the database that would support the application with the following attributes:

- Ability to support the high volume of write throughput placed on the database, which was expected to represent 50% of the total operational workload
- Deliver very low levels of latency and response times, in order to support the live, interactive experience expected by users of the service
- Achieve at least 99.999% uptime in order to meet the availability demands of users
- Scale linearly as the service grew, while at the same time, minimizing costs and complexity of both design and deployment

The documents managed by the DocQ application are stored on the Amazon S3 cloud-based storage service, with both the document's metadata and text from the actual pages extracted to xml and stored in the database. The pages are indexed by Sphinx, the open-source SQL full-text search engine. Document access is controlled by both the DocQ Access Control Lists (ACLs), maintained by the database, and Amazon's S3 ACL.
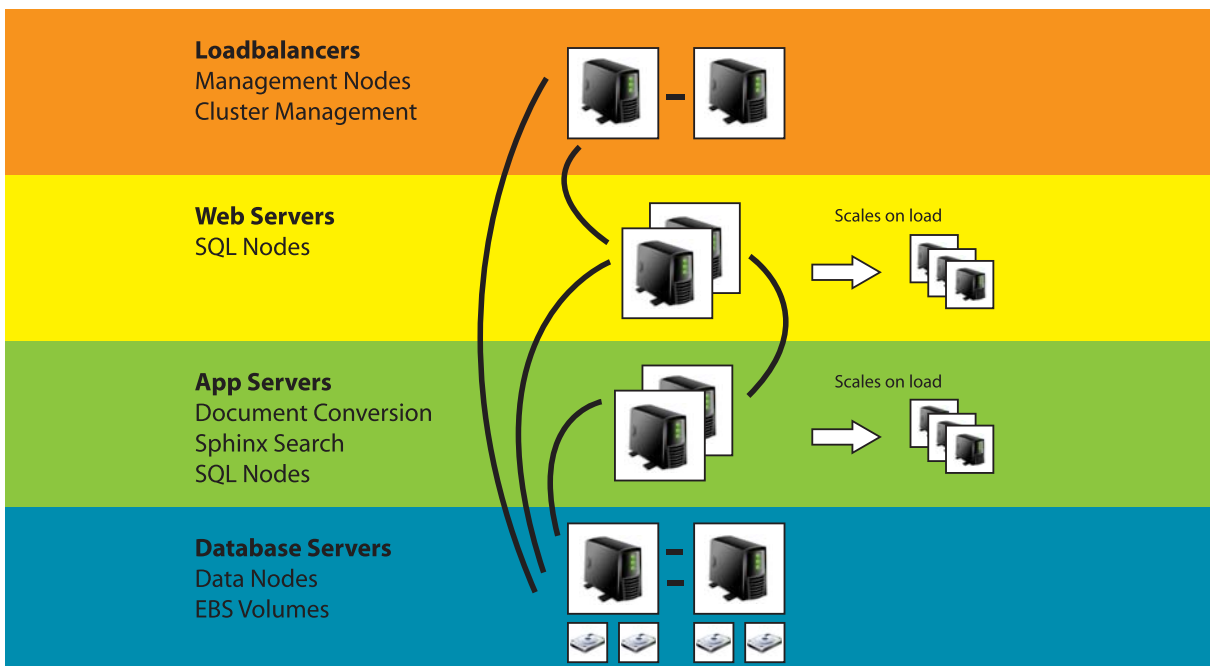
User session state is cached in the database to allow for service personalization, and then stored longer term in user tables from where older data is periodically purged.

"Because DocQ is first and foremost designed as a document transaction service, DocQ required a real-time architecture with millisecond level timestamp document identifiers. We were very excited when we found that MySQL Cluster could meet this requirement" said Casey Brown, DBA and Development Manager at Docudesk.

The MySQL Cluster Solution

In the past, Docudesk used MySQL to support some of their internal operations systems, and initially planned to use the MySQL Server with the InnoDB storage engine in a master / master sharded configuration to handle the write throughput requirements of the DocQ service. They also planned to deploy memcached in front of the MySQL servers in order to reduce database load and deliver lower latency read-access to users. However, the Docudesk architects were concerned at the level of complexity this type of solution would involve and decided to explore other options.

"For our update-intensive web service which demands 99.999% uptime, the MySQL and memcached solution would have been too complex. We found MySQL Cluster gave us the ability to scale our web services with low latency, right out of the box, in one solution" said Casey Brown, DBA and Development Manager at Docudesk.

Architectural Diagram of the DocQ Web Service

"For our update-intensive web service which demands 99.999% uptime, the MySQL and memcached solution would have been too complex. We found MySQL Cluster gave us the ability to scale our web services with low latency, right out of the box, in one solution."

Casey Brown
DBA and Development Manager, Docudesk

Through research, Docudesk found an academic paper describing the MySQL Cluster database, and realized that it offered the low latency, write-intensive support they needed, complemented by a fully redundant shared-nothing architecture needed to deliver 99.999% uptime. All of this functionality was available as part of the standard product offering, therefore eliminating the complexity of having to integrate multiple components into a single data management platform.

As a result of its open source licensing, Docudesk was able to freely download, evaluate and then deploy MySQL Cluster to power the DocQ service, without upfront licensing or support fees.

Since 2008 the infrastructure powering the DocQ application has been hosted on the Amazon EC2 and S3 cloud platforms, giving Docudesk on demand scaling. Currently, MySQL Cluster manages 4GB of data, which is growing at 1 to 2% every day, split between in-memory and disk-based tables. The MySQL Cluster configuration comprises two data nodes and two management nodes for redundancy and six MySQL Server application nodes providing SQL access to the data, handling an average of 500,000 to 1 million queries per day. Disk-based tables and Global Check Points (GCPs) are stored on Amazon Elastic Block Storage.

Text pages from the documents are stored as BLOBs (Binary Large Objects) in MySQL Cluster tables, allowing for fast text-string searches.

In addition to storing document metadata and text pages, PHP session data and ACLs, MySQL Cluster is also responsible for:

- maintaining the job queuing system (Amazon's own system did not deliver the low levels of latency or robustness demanded by the service)
- updating and recording all actions made to the documents (i.e. edits, annotations, addition of digital signatures, etc.) for use by the DocQ billing system.

The Future with MySQL

Docudesk plans to deploy MySQL Cluster 7.1 to take advantage of functionality enhancements offered by the latest release. They are also testing adding data nodes to the running cluster in order to scale database capacity and performance on-line, without interrupting service to the clients. Docudesk is also interested in the "push down joins" project that has the possibility of potentially improving multi-table joins if incorporated into future releases of MySQL Cluster.

Docudesk and MySQL Cluster Solution Overview

- The DocQ web service is a complete paperless business solution, providing a single repository for customers to manage, archive, and send their important documents.

- The database needed to deliver the high levels of write throughput, low latency responsiveness and continuous availability demanded by the service

- A sharded, multi-master MySQL solution with memcached was rejected due to the complexity of integration and management

- MySQL Cluster selected as it met all of the requirements of the service with one, integrated solution

- MySQL Cluster is handling 1 million queries per day on average across in-memory and disk-based tables, growing at up to 2% daily

- MySQL Cluster handles document metadata and text, PHP session state, ACLs, job queues and document actions for billing

MySQL Cluster

The Leading Open Source, High Availability Database for Real-Time, Mission Critical Applications

MySQL Cluster is the industry's only true real-time database that combines the flexibility of a high availability relational database with the low TCO of open source.

Carrier Grade Availability

MySQL Cluster features a "shared-nothing" distributed architecture with no single point of failure to assure 99.999% availability, allowing you to meet your most demanding mission-critical application requirements.

High Throughtput and Low Latency

MySQL Cluster's real time design delivers consistent, millisecond response times with the ability to service tens of thousands of transactions per second.
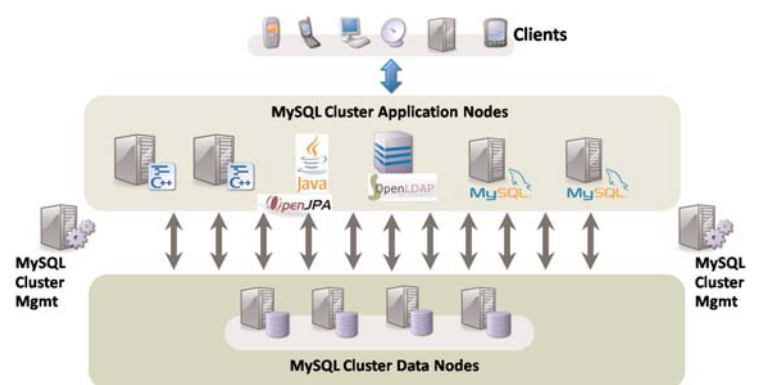
Linear Scalability

Support for disk based data, automatic data partitioning with load balancing and the ability to add nodes to a running cluster with zero downtime allows almost unlimited database scalability to handle the most unpredictable web-based workloads.

MySQL Cluster Target Applications:

- AAA / RADIUS / Diameter Data Stores
- Application Servers
- Data Store for LDAP Directories
- DNS/DHCP for Broadband
- eCommerce
- Mobile Content Delivery

- Online application stores and portals
- Payment Gateways
- Service Delivery Platforms
- Subscriber Databases
- VoIP, IPTV & Video on Demand
- Web Session Stores

The MySQL Cluster architecture has been designed for 99.999% availability and delivers massive read/write scalability

About MySQL

MySQL is the world's most popular open source database software. Many of the world's largest and fastest-growing organizations use MySQL to save time and money powering their high-volume Websites, critical business systems, communication networks, and commercial software.

For more information, go to **www.mysql.com/cluster**

ORACLE®