

## training - Bug #2459

### Archive Historical MongoDB Collection Data to S3 (CSV + Compressed Format)

04/07/2026 02:11 AM - Prashant Jain

<b>Status:</b>	New	<b>Start date:</b>	04/07/2026
<b>Priority:</b>	Normal	<b>Due date:</b>	
<b>Assignee:</b>	Shreya Agarwal	<b>% Done:</b>	0%
<b>Category:</b>		<b>Estimated time:</b>	0.00 hour
<b>Target version:</b>		<b>Spent time:</b>	0.00 hour
<b>Description</b>			
<p>We need to implement a process to archive historical data from MongoDB collections. The archival process should extract old data, convert it into CSV format, compress it, and store it securely in an S3 bucket for long-term storage and cost optimization.</p>			
<p>Scope of Work:</p>			
<p>Data Identification</p>			
<p>Identify historical data based on configurable criteria (e.g., records older than X months).</p>			
<p>Ensure active/critical data is excluded.</p>			
<p>Data Extraction</p>			
<p>Fetch data from MongoDB collections in batches to avoid performance impact.</p>			
<p>Support multi-tenant collections (if applicable).</p>			
<p>Data Transformation</p>			
<p>Convert extracted data into CSV format.</p>			
<p>Ensure proper schema mapping and column headers.</p>			
<p>Compression</p>			
<p>Compress CSV files using one of the following formats:</p>			
<p>.zip (preferred)</p>			
<p>.gzip (optional)</p>			
<p>Ensure optimal file size for storage and transfer.</p>			
<p>Storage (AWS S3)</p>			
<p>Upload compressed files to designated S3 bucket.</p>			
<p>Define folder structure:</p>			
<p>s3://&lt;bucket-name&gt;/&lt;tenant&gt;/&lt;collection&gt;/&lt;year&gt;/&lt;month&gt;/</p>			
<p>Enable versioning and lifecycle policies (if required).</p>			
<p>Post-Archival Handling</p>			
<p>Option to delete or mark archived records in MongoDB.</p>			
<p>Maintain audit logs for archived data.</p>			
<p>Automation</p>			
<p>Schedule via cron/job scheduler (e.g., daily/weekly).</p>			
<p>Retry mechanism for failed uploads.</p>			
<p>Acceptance Criteria:</p>			
<p>Historical data is correctly identified and extracted.</p>			
<p>Data is converted into valid CSV format.</p>			
<p>Files are compressed and uploaded to S3 successfully.</p>			
<p>No performance impact on live MongoDB operations.</p>			
<p>Logs are maintained for audit and traceability.</p>			
<p>Process is configurable (date range, collection name, tenant, etc.).</p>			
<p>Technical Considerations:</p>			
<p>Use streaming/batch processing to handle large datasets.</p>			
<p>Ensure secure S3 upload (IAM roles / access keys).</p>			
<p>Handle failures and partial uploads gracefully.</p>			
<p>Validate data integrity after archival.</p>			