International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Hadoop-to-Cloud Migration in Healthcare: Lessons and Case Studies

Selvakumar Kalyanasundaram

Texas, USA
inboxofselva@gmail.com

Abstract:
This paper explores the accelerating shift from traditional on-premises Hadoop clusters to modern, cloud-native data platforms, with a focus on Google Cloud. It highlights the limitations of Hadoop, including steep maintenance costs, performance bottlenecks, fragmented ecosystems, and lack of real-time processing capabilities. The study reviews key migration approaches—lift-and-shift, hybrid models, and full re-platforming—and examines the unique challenges enterprises face, such as large-scale data transfer, schema translation, compliance, and operational continuity. Through detailed case studies, including migrations by Seattle Children's Hospital and the Australian Department of Health and Aged Care, the paper illustrates how organizations have leveraged Google Cloud services like BigQuery, Dataproc, and Dataplex to achieve improved scalability, faster analytics, and reduced operational overhead. Lessons learned emphasize phased transitions, cloud-native modernization, governance, and the importance of training and change management. Ultimately, this paper demonstrates how cloud migration not only reduces infrastructure burdens but also enables innovation in advanced analytics and AI, particularly within healthcare.

Keywords: Google Cloud, Health care, migration, Hadoop, best practices.

INTRODUCTION

Companies are increasingly migrating from Hadoop platforms to the cloud due to scalability, flexibility, and cost-efficiency concerns. Traditional Hadoop clusters require significant on-premises hardware investment and ongoing maintenance, which can become expensive and difficult to manage as data volumes grow. Cloud platforms, on the other hand, offer on-demand scalability, allowing businesses to process massive amounts of data without worrying about physical infrastructure. Additionally, modern cloud-native tools such as BigQuery, Snowflake, and Databricks provide faster performance, serverless architecture, and simplified management compared to Hadoop's complex ecosystem. The cloud also enables seamless integration with advanced analytics, AI/ML services, and global collaboration features that Hadoop struggles to support efficiently. Furthermore, the shift reduces operational overhead, as cloud providers handle updates, security, and reliability, freeing enterprises to focus on deriving value from their data instead of maintaining infrastructure. As organizations push toward digital transformation and real-time analytics, cloud-based solutions present a more agile and future-ready alternative to legacy Hadoop systems.

LIMITATIONS OF HADOOP

Hadoop, despite being a pioneer in big data framework, has few disadvantages which lead many organizations to switch to cloud platform in recent years. Hadoop has a steep learning curve and requires specialized skill and heavy investment to install, configure, monitor clusters, enforce governance/security, manage HDFS health, schedule jobs, and to perform continuous patching and upgrading. Hadoop's core, i.e. MapReduce, was designed for mainly for batch processing, which means it is not suitable for real-time or low-latency applications. Other frameworks like Apache Spark or cloud-native solutions handle

streaming and interactive queries much more effectively. Hadoop Distributed File System (HDFS) relies heavily on disk-based storage, which makes it slower compared to modern in-memory or cloud-native storage systems. As data volumes increase, performance bottlenecks can become significant. Although Hadoop was initially considered cost-effective compared to traditional databases, the cost of maintaining large on-premises clusters (hardware, electricity, cooling, and administration) often outweighs the benefits.

Hadoop Cost

License / Support costs	$ 7,168 /Node/Year
Staffing	~4-8 nodes /100 nodes
Power and cooling	> $800 / Server / Year
Efficiency improvements	~10-40%

Sources: [1][2]

Hadoop's built-in security features (such as Kerberos) are complex and not user-friendly. It also lacks robust, easy-to-use data governance, auditing, and compliance capabilities compared to newer platforms. While Hadoop is technically scalable, scaling requires adding more physical machines, which becomes costly and harder to manage. Elastic cloud-native services scale much faster without infrastructure headaches. Hadoop's ecosystem (Pig, Hive, HBase, Oozie, etc.) is large but fragmented. Integrating and managing multiple components often leads to compatibility issues and higher maintenance overhead. With the rise of cloud-based data platforms and Spark, Hadoop is losing popularity, which means less innovation, fewer updates, and shrinking community support

MIGRATION APPROACHES

A. Lift-and-Shift Migration

Recreating the Hadoop environment on GCP with minimal changes (often as a first step). This typically uses Dataproc or even direct VM lift-and-shift. This approach speeds up migration and curbs data center costs quickly, though it may not immediately leverage cloud-native optimizations.

B. Hybrid Migration (Phased)

Running on-premises Hadoop and GCP in parallel during transition to avoid downtime. Large enterprises often replicate data incrementally to the cloud and migrate workloads in phases. Hybrid architectures ensure continuity for critical applications – data is synced between on-prem HDFS and cloud storage so both environments stay consistent. Healthcare Companies use this pattern to validate results on cloud while keeping on-prem systems live, then cut over when ready. A phased approach was recommended if a full cutover wasn't feasible immediately, enabling some workloads to remain on-prem temporarily while others move to GCP.

C. Re-platforming / Modernization

Redesigning the data architecture using cloud-native services. Rather than rebuilding Hadoop clusters on cloud, many organizations choose to re-platform onto managed services – for example, replacing Hive and HDFS with BigQuery and GCS, or converting batch ETL jobs to Dataflow. This pattern involves more transformation (e.g. rewriting SQL, changing data formats), but yields greater long-term benefits in performance and maintenance.

CHALLENGES IN MIGRATION

A. Data Volume

Hadoop clusters often host petabytes of data. Transferring such volumes over networks can be time-consuming and expensive. Migrating this require careful planning to avoid bottlenecks. Companies used

parallel transfer tools. (e.g. Hadoop DistCp with many mappers, or Google's Storage Transfer Service, Infoworks Replicator, Wandisco etc.,)

B. Schema and Code Translation

Hadoop ecosystems include varied tools (Hive, Pig, Spark, HBase, etc.) with custom schemas and code (HiveQL, MapReduce code, etc.). Migrating to GCP often means translating Hive schemas and SQL into BigQuery SQL dialect or converting ETL workflows to new tools. There are differences in data types and behavior (e.g. Hive custom data types vs. BigQuery or differences in partitioning). Google Cloud's BigQuery migration tools help by extracting Hive metadata and translating table definitions for BigQuery. . Automated translators (Google's built-in or third-party like Datametica - Raven) were commonly used to convert HiveQL, Spark jobs. Nowadays AI agents were preferred to convert the code and SQL (e.g. Vertex AI with LLMs).

C. Regulatory

Compliance and Security: Healthcare data is often subject to HIPAA and PHI regulations, and retail data may include sensitive customer PII and payment information (PCI). Migrating such data to cloud triggers concerns about security, privacy, and compliance. Enterprises must ensure cloud environments are configured to meet compliance – e.g. enabling encryption, policy tags, access controls, VPC isolation, and GCP's HIPAA compliance measures. Part of this involves re-evaluating user access privileges in the cloud, using tools like Identity and Access Management (IAM) to enforce least privilege.

D. Operational Continuity

Migrating mission-critical analytics without disrupting business operations is a delicate task. Companies often ran the old Hadoop jobs and new cloud jobs side-by-side on recent data to compare outputs. Automated data validation tools were adopted to speed this up (Datametica's Pelican, Collibra Owl, etc.,) was used in one case to perform cell-level validation between the source Hadoop data and target BigQuery tables. Ensuring disaster recovery and backup processes are maintained during migration is another concern (cloud architecture must provide at least the same level of DR as the on-prem Hadoop). Furthermore, performance tuning in the new environment is needed so that reports or batch processes run within expected timeframes. Some organizations reported initial cloud cost or performance surprises, underscoring the need to benchmark and optimize (e.g. adjust BigQuery slot commitments or Dataproc cluster sizing).

E. Data Format and Integration Issues

Hadoop data lakes often use formats like Avro, Parquet, ORC, etc. These generally work in GCP (BigQuery can directly ingest Parquet/ORC, and Dataproc/Dataplex handle them). Integration with existing systems can be challenging too. For example, healthcare providers might have EMR/EHR systems on-prem that feed data to Hadoop; during migration, pipelines had to be re-pointed to the cloud (perhaps via VPN or Transfer Appliances if direct cloud connectivity was limited). Streaming data ingestion (e.g. Kafka topics feeding Hadoop) had to be redirected to cloud equivalents like Pub/Sub or Kafka on GCP. Ensuring all upstream and downstream applications continue to function after the data moves is a considerable effort.

F. Skill Gaps and Change Management

Both sectors faced the human aspect of migration. Hadoop operations teams needed to learn GCP services, which required training. Companies also had to reorganize processes – for example, single large Hadoop cluster might has to be decentralized into many team-owned Dataproc clusters on GCP, which require new governance around cost attribution and cluster management. Change management is essential: teams must adapt workflows (e.g. using Airflow/Cloud Composer instead of Oozie or CRON, using GCS instead of

local HDFS). A successful migration often involved establishing a cloud Center of Excellence or partnering with experienced consultants to bridge knowledge gaps and follow best practices.

CASE STUDIES

To illustrate these trends, here are several real-world case studies from healthcare companies that migrated Hadoop to Google Cloud in recent years:

Seattle Children's Hospital (US) , a major pediatric hospital migrated 650+ databases and large datasets from on-prem systems (including Hadoop-based analytics) into BigQuery as their new enterprise data warehouse. The migration was part of a strategy to enable advanced analytics like NLP and medical image analysis with AI. By moving to BigQuery (paired with tools like Looker for BI), Seattle Children's achieved a ~30% lower TCO and built a sustainable growth platform for machine learning and data-driven care [10]. This case exemplifies re-platforming to a cloud data warehouse to meet healthcare analytics needs.

Department of Health and Aged Care (Australia) , a government health department began migrating its Hadoop-based data workloads into BigQuery as the core of a new enterprise data and analytics platform. The project's first phase moves HDFS data and Hive tables into BigQuery, consolidating data from various internal sources. Google Cloud noted that centralizing data in BigQuery will improve compliance with data regulations and yield substantial cost and efficiency benefits for the department's analytics programs. Importantly, the plan includes training dozens of staff on the new platform to ensure a smooth transition. [11] This case highlights a public healthcare entity using a cloud data warehouse approach (BigQuery) to replace on-prem Hadoop, emphasizing compliance and scalability.

Overall, healthcare firms migrating to GCP often focus on enabling secure data sharing, faster research analytics, and cutting IT maintenance costs.

Etsy, an online marketplace, decided in 2017 to move off its on-premises Hadoop infrastructure and fully embrace Google Cloud. The migration was completed by 2019, taking about two years. According to Etsy's CTO, this shift to GCP paid off by redistributing 15% of engineering staff away from routine infrastructure maintenance toward customer-facing improvements. [3]. In a Google case study, Etsy reported the cloud transition met key goals: it ensured smooth handling of holiday traffic spikes, improved cost-efficiency, and even reduced the company's environmental footprint. In short, Etsy's Hadoop-to-GCP migration enabled faster innovation and reliable scalability. Twitter operates one of the world's largest Hadoop data platforms on-prem. In 2018, Twitter announced a partnership with Google Cloud to migrate large portions of its data backend to GCP. This involves moving over 300 petabytes of Hadoop files into Google's cloud. The expected outcome is a more agile data infrastructure with improved disaster recovery and security, supporting Twitter's growth [4].

One of the National U.S. healthcare system , migrated from both Cloudera and Oracle to Google Cloud. With Google Cloud, users can act on insights much faster, especially when tracking community health. They were able make daily (not monthly) data refreshes, response times that are 25–50× quicker than the previous Oracle system, and a unified dashboard interface that makes information easier to find [5].

On the other side of the trend, some U.S. companies that ventured into the cloud have pulled back and re-invested in on-premises infrastructure, including Hadoop or similar big-data platforms. These reversals (often called "cloud repatriation") usually cite uncontrolled costs, performance issues, or regulatory concerns. In effect, these firms decided that running certain analytics and data workloads in their own data centers (often leveraging open-source tools like Hadoop, Spark, etc.) was better than continuing in public cloud.

GEICO's decade-long effort to modernize on Azure cloud largely failed to meet its objectives. The company discovered that moving legacy systems to cloud without redesign ("lift-and-shift") resulted in higher costs and complexity rather than simplification. In GEICO's case, the cloud migration introduced availability issues (more downtime) and no unified data strategy, all while the cost of running IT went well above on-prem levels.[6] . In Capital One, a misconfiguration in an AWS S3 storage bucket left a huge trove of customer data exposed on the cloud. It underlined the importance of rigorous cloud security practices and shared responsibility. Beyond these examples, surveys show many enterprises have had to adjust course mid-migration. Common problems include budget overruns. In fact, a 2019 study by Fortinet found 74% of companies moved some applications back on-prem after cloud migrations failed to deliver anticipated benefits [8]. 65% of organizations have made strategic investments in cloud, but only 32% are achieving their ambitions.[9]. These cases highlight the varied outcomes – from successful modernization to costly rollback – that enterprises have experienced in migrating big data from Hadoop to the cloud. Each provides valuable lessons for future migration strategies.

Over the past three years, many enterprises in the healthcare and retail sectors have migrated their on-premises Apache Hadoop data platforms to Google Cloud. Organizations are moving away from aging, costly Hadoop clusters toward managed services like Google Cloud's Dataproc and BigQuery. Companies can opt for one of the below migration patterns based on their business objectives and desired outcomes. these patterns are not mutually exclusive – many migrations start as lift-and-shift, then evolve into full re-platforming

LESSONS LEARNED AND BEST PRACTICES

Across these migrations, enterprises have gleaned important lessons and established best practices to ensure success:

A. Thorough Assessment & Benchmarking

Before migration, measure your current Hadoop cluster's performance and costs, then pilot on GCP to compare. Understand workload characteristics (CPU, memory, I/O patterns) so you can right-size cloud resources. Cost modeling is crucial: on-prem costs are fixed, whereas cloud costs scale with usage. Understand the key job behaviors and do the financial analysis of running them in cloud vs on-prem.

B. Plan for a Phased Transition

Migrate incrementally, focusing on one domain or workload at a time. Many firms first tackled storage migration (offloading HDFS to GCS) and later moved compute workloads. During migration, run both systems in parallel for a period to validate results and stability. Prioritize migrating stable, less critical workloads initially as a learning experience, then move mission-critical jobs once the team is confident.

C. Leverage Cloud-Native Services

Take the opportunity to modernize using cloud services (BigQuery, Dataflow, Pub/Sub, etc.) instead of simply cloning the Hadoop stack on VMs. This yields better long-term outcomes (serverless auto-scaling, less ops burden, etc.). Establish a partnership with cloud providers for any edge use-cases to mitigate risk.

D. Ensure Governance, Security & Compliance

Set up Cloud IAM roles and resource hierarchies so that each team/project has proper data isolation. Use encryption (Cloud KMS) for sensitive data. Enable audit logs and possibly VPC Service Controls to prevent data exfiltration in healthcare scenarios. Define data catalog, quality checks, PII masking wherever needed. Prefer Dataplex for organizing and governing the data across distributed storage.

E. Optimize Data Architecture (Decouple Storage/Compute)

Cloud migration is a chance to redesign for efficiency. One best practice is separating storage from compute – e.g., storing persistent data in GCS or BigQuery instead of tied to compute nodes. This was done by most organizations (GCS as a drop-in for HDFS). It allows compute clusters to be ephemeral and auto scaled, reducing costs . Unifying data lake and warehouse can avoid redundant pipelines. for instance, using BigQuery external tables on Parquet files or leveraging BigQuery for both ETL and BI. Also,

consider data partitioning and clustering in BigQuery to optimize query performance on large migrated datasets.

F. Invest in Automation (Migration Utilities & Testing)

Use available migration utilities: e.g., Google's BigQuery batch SQL translator to convert Hive/Oracle SQL to BigQuery SQL, or schema conversion tools to rebuild Hive tables on BigQuery. For Hadoop file migrations, tools like gsutil rsync or DistCp with GCS connector can automate copying large directories. Moreover, automate data validation and pipeline testing. Automation also applies to infrastructure-as-code: define your GCP infrastructure (networks, IAM, Dataproc clusters, BigQuery datasets) in code and deploy repeatably, so you can confidently recreate environments and avoid manual config errors.

G. Mind Cloud Quotas and Limits

When planning big migrations, account for cloud quotas (compute, storage, IPs, etc.). It's best to engage Google Cloud support early to raise quotas in target regions and to design around any limits (e.g., BigQuery load jobs per day, or API request rates). Similarly, take advantage of cost controls like budgets and alerts during the migration.

H. Use Cost-Optimization Features Wisely

Cloud offers new levers for cost savings, which should be used from the start. For instance, taking advantage of preemptible VMs for Dataproc (for non-critical, fault-tolerant jobs) can drastically cut compute costs. Right size your cluster nodes or BigQuery slot capacity based on benchmarks. Also, consider scheduling jobs to run in cost-effective ways (e.g., batch jobs during off-peak hours, using auto-scaling clusters to only pay for what you need). In summary, continuously optimize: once workloads are running in GCP, review usage patterns and tune resource allocations or reservations to save money. Set up monitoring process to identify Excessive BigQuery Scans and Unoptimized Queries. In BigQuery, always partition large tables (typically by date or another natural partition key like hospital ID or year) and cluster by frequent filter columns

I. Avoid Storage Sprawl and Data Redundancy

Cloud storage is cheap per GB, but at multi-terabyte scales, costs and management burdens add up. Storage sprawl happens when there is no discipline in organizing and retiring data. Healthcare organizations are especially prone to this if governance is siloed; a research team might clone a dataset rather than reuse a curated version, leading to multiple silos of the same PHI data. Additionally, storing everything forever (due to fear of deleting potentially useful data) means old or unused data piles up. Over years, this cold data can become a significant cost center if not archived to cheaper storage or purged. High storage footprints also slow down operations (e.g. longer backups, larger metadata catalogs to maintain). Develop a clear data lifecycle policy to prevent the uncontrolled growth of data and ensure quality over quantity.

J. Train Teams and Update Processes

Provide hands-on training for data engineers, analysts, and admins on GCP tools (BigQuery, Dataproc, GCS, etc.). A best practice is to start a Cloud Center of Excellence or at least designate cloud champions in each team, who can disseminate best practices and act as liaisons with cloud experts. Additionally, engage a trusted partner or Google's Professional Services if your team lacks specific.

K. Focus on One Objective at a Time

Migrations are complex. It is better to set clear phased goals. for example.
Phase 1. Move and replicate data to cloud storage with zero downtime
Phase 2. Get critical ETLs running in Dataproc with same output
Phase 3. Migrate Hive warehouses to BigQuery and decommission Hadoop.
Phase 4. Focus on cost optimizations and performance tuning

MODERN CLOUD DATA ARCHITECTURE

A. Zoned Architecture

Healthcare organizations that migrate off Hadoop typically adopt a modern data architecture on Google Cloud that leverages scalable storage and serverless compute. Google Cloud Storage (GCS) often serves

as the data lake for raw and unstructured files, while BigQuery becomes the core analytical data warehouse. In this architecture, raw data from on-premise Hadoop/HDFS is moved into Cloud Storage buckets or directly into BigQuery tables, and Apache Hadoop processing jobs are refactored using Google's managed services (e.g. Dataflow streaming pipelines or Dataproc for any Spark/Hadoop jobs). This provides a clear separation of storage and compute: Cloud Storage handles durable object storage of datasets (e.g. CSV, JSON, DICOM images), and BigQuery provides a scalable SQL engine for analysis of structured data. The compute layer is largely serverless. BigQuery itself handles massively parallel query execution, and tools like Dataproc or Dataflow can be used on-demand for ETL or streaming, eliminating the need to manage Hadoop clusters. This cloud-native architecture improves scalability and performance. Healthcare data platforms in GCP are organized into tiered zones

1. Bronze zone: It is the Raw layer, the landing area for raw, unprocessed data from various sources. Bronze zone act as a staging area and system of record of the incoming data before any cleaning.
2. Silver zone: A middle layer where data is cleansed, standardized, and integrated for consistency. In this zone, healthcare organizations perform transformations such as mapping HL7 messages to FHIR resources, normalizing or de-duplicating patient records, standardizing codes (ICD, CPT, LOINC), and applying data quality checks. The curated data is often stored in structured formats (Parquet/Avro in GCS or refined BigQuery tables) with enforceable schemas. This zone ensures the data is analytics-ready and interoperable.
3. Gold zone: The final business-ready data layer optimized for analytics, reporting, and machine learning. In healthcare, this "gold" zone contains curated data reshaped into summary tables for specific use cases This gold zone is what downstream analysts, data scientists, or applications query directly. It is often subject to strict access controls and may incorporate row-level security or data de-identification where appropriate to protect patient privacy. The end result is a set of trusted, high-quality datasets that drive dashboards, analytics tools, and ML models.

This zoned architecture is valuable in healthcare because it improves data quality and governance at each step. By separating raw data from refined data, organizations can maintain a clear audit trail and ensure regulatory compliance (e.g. HIPAA) during data transformations. Google Cloud's Dataplex service can further facilitate this zoning: Dataplex lets you define data lakes and subdivide them into zones (e.g. a raw zone and curated zone) with managed metadata and policy enforcement. Each zone in Dataplex can encompass multiple storage systems (GCS buckets or BigQuery datasets) yet be managed under a consistent governance framework. In practice, a healthcare provider might create a Dataplex lake for each major domain (e.g. a "Clinical" lake and a "Claims" lake), and within each lake, have a raw zone for landing new data and a curated zone for the cleaned, standardized data ready for analytics. This approach enforces a clear separation between raw and curated data while simplifying discovery and policy management.

B. Domain-Based Organization of Data

Healthcare data is extremely broad, so organizing datasets by domain is a common practice after migrating to Google Cloud. Rather than a monolithic data swamp, organizations logically group data by subject (like pharmacy, prescriber, prescription, claims, etc.,). Each domain can correspond to its own BigQuery dataset. This improves the quality and governance of the data. Similarly, separate the development, test, and production data to avoid accidental use of test data in production analyses. It helps to avoid security and compliance risks.

CONCLUSION

The migration of big data platforms from Hadoop to Google Cloud reflects a broader enterprise drive toward agility, scalability, and digital transformation. While Hadoop pioneered distributed data processing, its high maintenance requirements, limited real-time capabilities, and fragmented ecosystem have become

barriers in today's fast-paced analytics landscape. Google Cloud provides a compelling alternative, offering serverless, elastic, and integrated solutions that empower organizations to modernize their data architecture. However, the transition is not without challenges—enterprises must carefully manage data volumes, compliance risks, and skill gaps to avoid costly missteps. Successful migrations show that phased approaches, cloud-native re-platforming, and strong governance frameworks yield the best outcomes. For healthcare in particular, cloud adoption supports secure data sharing, advanced analytics, and AI-driven care delivery. Looking forward, enterprises that embrace these lessons and optimize their cloud strategies will be well-positioned to unlock long-term value from their data, transforming not just their IT operations but also their capacity to innovate.

REFERENCES:

1. The Total Economic Impact™ of The Dell | Cloudera Apache Hadoop Solution, Accelerated By Intel Cost Savings And Business Benefits Enabled By The Dell | Cloudera Apache Hadoop Solution, Accelerated By Intel [2015]
2. https://www.databricks.com/blog/2021/03/25/its-time-to-re-evaluate-your-relationship-with-hadoop.html
3. https://www.ecommercebytes.com/C/blog/blog.pl?/pl/2020/11/1605715268.html
4. https://blog.x.com/engineering/en_us/topics/infrastructure/2018/a-new-collaboration-with-google-cloud#:~:text=CTO%20of%20Google%20Cloud
5. https://resources.pythian.com/hubfs/Case-Study/Healthcare-System-Transforms-Its-Data-Infrastructure.pdf
6. https://www.thestack.technology/warren-buffetts-geico-repatriates-work-from-the-cloud-continues-ambitious-infrastructure-overhaul
7. https://www.cloudcomputing-news.net/news/10-real-life-cloud-security-failures-and-what-we-can-learn-from-them
8. The Bi-Directional Cloud Highway: User Attitudes about Securing Hybrid- and Multi-Cloud Environments , 2019 Jeff Wilson Senior Research Director, Cybersecurity Technology
9. https://www.hfsresearch.com/research/only-a-third-of-enterprises-are-realizing-their-cloud-ambitions/
10. New incentives and tools to migrate to BigQuery and Dataproc | Google Cloud Blog
11. https://www.itnews.com.au/news/department-of-health-and-aged-care-to-deploy-enterprise-data-platform-605522