

# HADOOP TO SNOWFLAKE MIGRATION

## Whitepaper



## TABLE OF CONTENTS

Why make a change? .....	3
Why Snowflake? .....	3
Snowflake Considerations .....	3
Meet phData.....	4
A Snowflake Partner to Get it Done Right.....	4
Automate Your Migration (and more!).....	4
Data Lake vs. Data Warehouse .....	5
Data Lakes.....	5
Data Warehouses .....	5
Is our workload the right fit?.....	6
Snowflake and Spark.....	6
Hadoop to Snowflake Technology Mapping.....	7
Migration Plan .....	8
Top Level Strategy .....	8
Phase 1 .....	9
Tools and Technologies.....	9
Data Sources.....	9
Use Cases for Hadoop.....	11
End-User Training .....	11
Integrations.....	11
Security and Governance .....	11

Phase 2 .....	12
Storage Migration .....	12
Phase 3 .....	13
Data Validation.....	13
Case Study.....	14
The Customer .....	14
The Challenge .....	14
The Strategy.....	14
The Outcome.....	14
Getting Started.....	15

### PURPOSE

*This document is intended to serve as a general roadmap for migrating existing Hadoop environments — including the Cloudera, Hortonworks, and MapR Hadoop distributions — to the Snowflake platform. Each distribution contains an ecosystem of tools and technologies that will need careful analysis and expertise to determine the appropriate mapping of technologies that will ultimately best serve any given use case.*

# WHY MAKE A CHANGE?

The Hadoop platform represents a broad suite of technology offerings, requiring architects and engineers to select the right tool for the job. Because of this, businesses see a limited ROI from what they can derive from their data due to poor query performance, difficult to use tools, and bulky execution engines in Hadoop.

Managing both on-premises and cloud-based Hadoop clusters requires a dedicated infrastructure administration team to handle upgrades, security patches, capacity planning, and more.

That means successfully using Hadoop requires deeply technical users and administrators which are hard to find and highly expensive.

That's why Hadoop users are moving to Snowflake.

## WHY SNOWFLAKE?

The Snowflake Data Cloud was designed with the cloud in mind, and allows its users to interface with the software without having to worry about the infrastructure it runs on or how to install it. Between the reduction in operational complexity, the pay-for-what-you-use pricing model, and the ability to isolate compute workloads there are numerous ways to reduce costs associated with performing analytical tasks. Some other benefits and capabilities include:

- **Data sharing:** Easily share data securely within your organization or externally with your customers.
- **Zero copy cloning:** Create multiple 'copies' of tables, schemas, or databases without actually copying the data. This saves on the time to copy and reduces data storage costs.
- **Separate compute and storage:** Scale your compute and storage independent of one another, and isolate compute power for jobs that need their own dedicated warehouse.
- **No hardware provisioning:** No hardware to provision, just a t-shirt sized warehouse available as needed within seconds.



## SNOWFLAKE CONSIDERATIONS

Snowflake is built on public cloud infrastructure, and can be deployed to Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP). When moving to Snowflake, there are some considerations regarding which cloud platform to use. Refer to the [Snowflake documentation](#) to assist with choosing the right platform for your organization. For the purposes of this migration plan, AWS technologies will be used when options are available.

---

## PHDATA, YOUR PREMIER PARTNER

phData is a Premier Service Partner and Snowflake's Emerging Partner of the Year in 2020. If you're looking to migrate, phData has the people, experience, and best practices to get it done right. We've completed nearly 1,000 data engineering and machine learning projects for our customers. So whether you are looking for architecture, strategy, tooling, automation recommendations, or execution, we're here to help!

# MEET PHDATA

phData is a services company dedicated to delivering solutions on Snowflake. We help with end-to-end services to architect, deploy, and support machine learning and data analytics on Snowflake.

## phData Services

- **Data Engineering:** We bring proven tooling, automation, and engineers with deep Snowflake expertise to get your streaming, batch, and interactive data products into production.
- **Cloud DataOps:** We deliver 24x7 monitoring, management, and administration for your data pipelines and applications, with a focus on driving down costs and delivering the best user experience.
- **Machine Learning:** With services across the full ML lifecycle, our experienced data scientists and ML engineers help you build, train, and deploy ML models, then ensure those models continue delivering value.

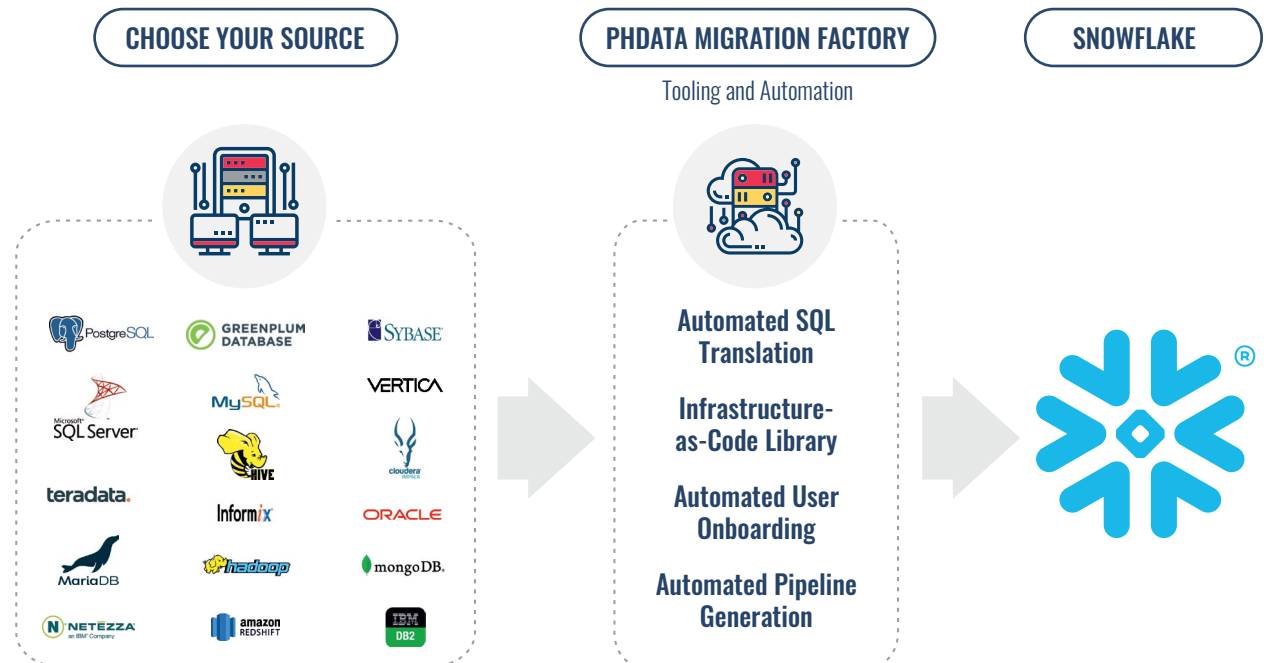
## A SNOWFLAKE PARTNER TO GET IT DONE RIGHT

phData has the people, experience, and best practices to get the job done right. By using our leaner, more specialized, and experienced team — as well as our software and automation — our customers typically save 20–35 percent on their pipeline implementation and support costs.



## AUTOMATE YOUR MIGRATION (AND MORE!)

Whether you're migrating your existing on-premises data warehouse to Snowflake or building an all-new cloud-native solution, our Snowflake specialists help you break down data silos faster.



# DATA LAKE VS. DATA WAREHOUSE

Snowflake is both a data lake and a cloud data warehouse for semi-structured and structured data. These are not entirely separate technology frameworks in direct competition with one another. A data lake generally contains the raw data required to build a data warehouse.

## DATA LAKES

Data lakes have risen in popularity with the rise of big data technologies like Hadoop. Most Hadoop platforms start with a specific use case involving data warehousing already in mind; however, to actually accomplish that use case, the Hadoop platform turns into a data lake.



This is because data gets ingested into Hadoop from heterogeneous sources in many different formats, including structured, semi-structured, and unstructured data. Semi-structured and unstructured data often need to be reformatted in order to be consumed by a data warehouse.

### Features

- Data is often stored in its RAW format – structured, semi-structured, or unstructured; from any source
- Data can be stored with minimal processing and is only transformed when put to use

## DATA WAREHOUSES

Data Warehouses date back to the 1970s, and they became the norm in the late 1990's. Originally implemented in Relational Databases, the data is tabular, easy to query, and allows reports to be built on top of it.



### Features

- Structured, cleaned, processed, and refined data; typically from operational systems
- Data must be cleaning and refined before it is stored
- Data is transformed into a format which Data Scientists can tap into to build models and data applications that ultimately deliver business value through reporting and analytics

## BEST PRACTICE

Snowflake is both a cloud data warehouse and data lake which is ideal for semi-structured and structured data. phData's recommendation for truly unstructured data (e.g. videos, pictures) is to store it in the cloud object storage and then analyze it in Snowflake via an external read from the object storage layer. This allows you to make sense of the raw data and transform it into something a data warehouse can analyze and use.

# IS OUR WORKLOAD THE RIGHT FIT?

Nine out of ten of phData's customers use Hadoop for analytical workloads where the sources are mainly relational or tabular data, such as spreadsheet or delimited data. These analytical workloads are perfect targets to migrate to the Snowflake platform. Hive and Impala scripts can be easily migrated to Snowflake, and can be run either in Worksheets or via SnowSQL from the command line interface. phData also offers custom-built software to automate the translation of SQL dialects from Impala to Snowflake SQL.

## SNOWFLAKE AND SPARK

Snowflake offers various connectors between Snowflake and third-party tools and languages. One of these is a Spark Connector, which allows Spark applications to read from Snowflake into a DataFrame, or to write the contents of a DataFrame to a table within Snowflake.

Please note that Apache Spark applications will need to be reviewed and fully understood before migrations can occur. If SparkSQL is involved, the SQL code will be able to run in Snowflake. phData engineers are trained in converting apps like these; however, appropriate A/B testing will still need to be conducted.

On the other hand, online use cases that require lookups by key in sub-second might instead require another tool in the cloud ecosystem, depending on the customer's choice of vendor.

HBase and MapR DB, for example, are both distributed column-oriented databases built on top of the Hadoop file system. HBase serves a similar purpose as MapR DB, and has a very similar migration path. Use cases that are primarily OLTP driven will be converted to streaming applications utilizing Azure CosmosDB, DynamoDB or Google BigQuery for storage, while analytics applications can be migrated to a streaming application into Snowflake.

Snowflake announced a new search feature that allows quick lookups for given keys, which will eventually be able to replace the aforementioned technologies. Until then, phData can assist with a migration from HBase/MapR DB to CosmosDB/DynamoDB/BigQuery.



## HADOOP TO SNOWFLAKE TECHNOLOGY MAPPING









HADOOP TECHNOLOGY	SNOWFLAKE NATIVE	RISK	VERIFICATION	CLOUD TECHNOLOGY
HBase	Consult phData; Investigate Snowflake Search	High	Will it meet SLAs?	DynamoDB
				CosmosDB
				Google BigQuery
Hive	Yes, w/ SQL Transformation	Low		Snowflake
Impala	Yes, w/ SQL Transformation	Low		Snowflake
Kafka	Yes	Medium	Kafka Connect w/variant fields	Apache Kafka
				Kinesis
				Azure Event Hubs
Kudu + Impala	Yes, w/ SQL Transformation	Low		Snowflake
Solr	Consult phData; Investigate Snowflake Search	High		ElasticSearch
				Azure Search
				Dataproc Solr
Spark	No; but Snowflake Connector exists	Medium	Dependant on the workload	AWS EMR, AWS Glue
				Azure DataFlow
				GCP DataProc
SparkSQL	Streams and Tasks with SQL Transformation	Low		AWS EMR, AWS Glue
				Azure DataFlow
				GCP DataProc

# MIGRATION PLAN

The goal of migrating from Hadoop distributions to Snowflake is to offer you a straightforward path to becoming cloud-native and to save on licensing and hardware costs. Ultimately, migration plans and timelines can vary significantly, depending on the size of your organization and the complexity of your use cases.

## TOP LEVEL STRATEGY

Plan and execute your migration over three distinct phases: Discovery, Implementation, and Validation.

PHASE 1: DISCOVERY	PHASE 2: IMPLEMENTATION	PHASE 3: VALIDATION
 <b>Learn</b> Catalog and understand different types of workloads running in their cluster	 <b>Implement</b> Cloud-native architects and engineers will construct your foundational platform	 <b>Verify</b> Data and workloads will be reviewed and verified before cutover activities
 <b>Design</b> Size and design the new architecture that supports both data and applications	 <b>Move</b> Leveraging our experience and tooling, data and workloads will be migrated	 <b>Support</b> New projects, users, and any remaining workloads
 <b>Plan</b> Establish a detailed migration plan, ensuring disruptions are avoided		 <b>Evolve</b> Establish future roadmap, including ideas for the next phase of cloud



# PHASE 1 DISCOVERY

The goal of Discovery is to gather necessary context and background information about your current Hadoop environment, and to identify all the relevant dependencies. This includes tools and technologies, data sources, use cases, resources, integrations, and service level agreements. The outputs from this phase will be critical for informing the final migration plan.

## TOOLS AND TECHNOLOGIES

Creating a complete inventory of all tools and technologies used in your current Hadoop environment is critical to creating a successful migration plan. This inventory should include tools native to the Hadoop environment, as well as any relevant third-party vendor software.

From this inventory, phData will create a current-state architecture of your existing Hadoop landscape. Then, together we will identify whether each tool is still needed in the go-forward strategy, or whether it can be replaced by the capabilities of Snowflake. From there, those tools that will be retained have to be evaluated to fully understand how they work in a cloud-native Snowflake environment.

This technology inventory should be categorized as follows:

- Ingest and data integration
  - Storage formats
- Data processing and transformation
  - Structured vs unstructured
- Streaming
  - MapR streams or Apache Kafka
- Data cataloging, discovery, and lineage
  - Message format
- SQL interfaces
  - Volume
  - Source and target integrations
- Third-Party vendor software
- Security and governance
- Data visualization
- Storage

## DATA SOURCES

The actual data sources are what deliver results and provide value. That makes them just as important as the tools and technology inventory.

These sources may be external — such as Databases, ERP, CRM, files, streaming or event data — or they may be internal sources used for integrations feeding data out of Hadoop.

The discovery phase is an ideal time to identify any data sets that might be deprecated and don't need to move to the cloud. You should identify each data source to ensure it has

an owner or subject matter expert who can be available to answer questions and provide a detailed understanding of data sets.

phData will work with these data experts to complete the following list of questions:

### General Data Source Questions

- Who is the data owner/steward?
- What is the classification of data?
- What are the retention policies?
- Are there obfuscation or data masking requirements?
- Are there data encryption requirements?

### Relational Database Questions

- Database type (e.g. Oracle, Microsoft SQL server, DB2, MySQL, etc.)
- Are there any custom functions/UDFs that need to be migrated/replicated?
- Number of databases
  - Number of tables
    - Number of columns
    - Data types (specifically custom data types)
- Required frequency of ingest
- Bulk ingest or incremental
  - Can CDC operations be queried via audit logs?
  - Do non-CDC sources have a “last modified” or “last updated” column?
  - How are hard deletes handled in non-CDC sources?
- Do the table structures change frequently, adding or removing columns, data type changes, etc.?
- Security requirements, roles, users, downstream consumers

### MapR Streams or Kafka Questions

- What is the message type (e.g. AVRO, JSON, XML, delimited text, etc.)?
- Is there a schema registry?
- Number of topics
  - Number of partitions
    - Partition strategy, what are your message keys
  - Individual message size
  - Messages per second
  - Consuming applications
  - Data retention requirements

### Delimited File Questions

- Where are the files located?
- Are they accessible from the chosen cloud provider?
- Does new data get delivered via a new file or append to the existing file?
- What is the File format?
- What is the Delimiter format?
- Does the file contain record headers?
- Can schema be inferred on read?
- What is the File encoding?

### API Integration Questions

- How can programmatic access to the API be provided?
- What is the schema for responses?
- Is there a schema registry?
- Type of data?
- How many requests per second?
- Structure of requests (i.e. is the response a bulk load of a dataset)

or incremental change?)

- Does the response schema change over time?
- Authentication and authorization mechanisms?

## USE CASES FOR HADOOP

The most important step of the discovery phase is to fully understand the applications running in your current Hadoop environment. This will better define the amount of effort it will take to migrate your environment to the cloud.

Because Hadoop comprises a variety of tools and services that can make up an application, there are some important distinctions to make about how the application is deployed. phData will work with you and each application owner to understand the following:

- A complete description of the application and its value
- What technologies does the application use in the Hadoop environment?
  - Versions of these technologies e.g. Spark 1 vs Spark 2
- Is the application streaming or batch?
- What external libraries are used?
- Reference architectures
- Complete list of data sources
- Expected outputs
- Downstream consuming applications
- Security and Classification consuming and produced datasets
- Production readiness
  - Expected errors
  - Frequency of errors
  - Steps to resolution
  - Support and monitoring implementations

## END-USER TRAINING

Internal stakeholders and existing end-users tend to be concerned about change. Identifying the users of the current platform and providing them with the necessary training around Snowflake is crucial to accelerating platform adoption. Understanding tools and processes in their daily workflow and making alternatives available will make sure that these people feel comfortable using Snowflake.

## INTEGRATIONS

It's important to inventory applications accessing the Hadoop environment in order to ensure that the cloud-native Snowflake environment can serve data to them. In many cases, these applications have complex access patterns for authentication and authorization which will need to be evaluated. However, external applications connecting using JDBC or ODBC can use these drivers when communicating with Snowflake, and should work out of the box.

As mentioned before, there are also Connectors available to integrate with Python, Spark, Kafka, and more.

## SECURITY AND GOVERNANCE

Finally, you need to conduct an inventory of current security and governance configurations. Because the different architectures that each Hadoop distribution puts forth, the scope could broaden to include more tools, such as Sentry, Ranger, and HDFS ACLs. In addition, the use of Kerberos, Active Directory, encryption-at-rest and encryption-in-transit must all be taken into account. And for MapR, filesystem permissions, and ACLs, Access Control Expressions must be assessed.

phData has developed an automation tool called [Tram](#), which generates databases, schemas, roles and security grants in Snowflake. Tram is able to integrate with on-premises Active Directory as well as Azure AD. Additionally, it can also create users directly in Snowflake, and this latter method can be used within a git workflow.

# PHASE 2 IMPLEMENTATION

The implementation phase focuses on migrating business applications into Snowflake from your existing Hadoop environment. Based on the inputs gathered during the discovery phase, a list of data sources, applications, and tools will be selected and prioritized for migration.

From there, phData will provide a team of architects and data engineers to work with the customer throughout the migration effort. This phase will be the longest and most technical to execute.

## STORAGE MIGRATION

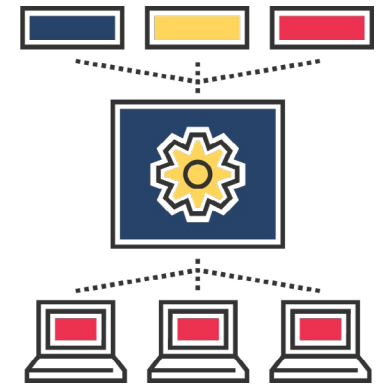
Using the Data Source inventory, engineers will move data stored in HDFS to the cloud vendor's storage layer (Blob Storage or S3). The same information architecture will be applied in the new cloud storage file system, and the resulting folder and file structure should be a one-to-one match with that of the HDFS file system.

Since use cases and implementations vary greatly from customer to customer, phData does not see any generalized quick-win migration tool as being feasible for the implementation phase. However, our engineers do build code and tooling to be used between migrations as patterns develop and new opportunities arise. We then provide it to the broader Snowflake community.

For example, we've developed a tool called SQLMorph to instantly translate Hadoop SQL to Snowflake SQL, which eliminates a usually time-consuming, error-prone, and highly manual process.

### Challenges

A significant challenge to the storage migration for both Cloudera and MapR implementations will be migrating role-based access controls to the new cloud storage system. Depending on the cloud vendor, the configuration of role-based access is different. Azure Blob Storage uses Active Directory, with groups and users to grant access to blob containers, whereas Amazon AWS uses IAM policies to configure access to S3 buckets. Accordingly, a tool will need to be developed to migrate these policies for each cloud vendor.



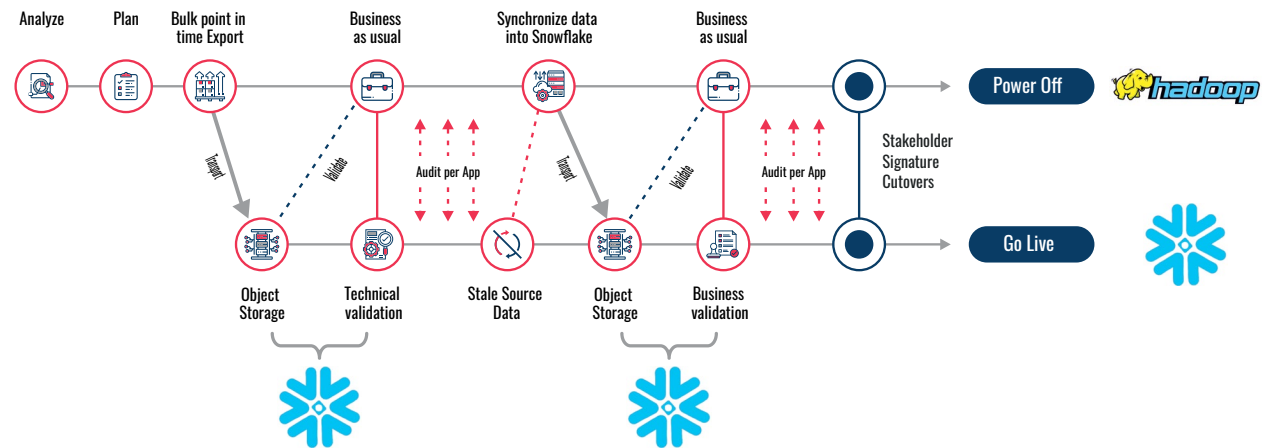
# PHASE 3 VALIDATION

The final phase will be to validate the outcome of the migration from Hadoop to Snowflake. This step should be performed using traditional A/B testing. For a time, you will need to continue running the existing Hadoop implementation alongside the new cloud-native Snowflake offering. Meanwhile, validation scripts and processes will be developed to ensure the delivered results in Hadoop match with those in Snowflake. After all checkouts have been performed and applications and use cases are cleared, they can be shut down in Hadoop.

After this phase, you should be able to completely remove or repurpose your Hadoop infrastructure.

## DATA VALIDATION

This diagram outlines our validation process:



The first validation point is ensuring the data in your existing Hadoop environment matches the data in your new Snowflake environment. To do this, we will run a query in Hadoop, run the same query in Snowflake, and ensure the data is an exact match. We use a suite of tools to automate this process.

Once the data is validated, we move to our technical validation phase to ensure the Snowflake environment has the required integrations that your development teams expect. This includes checking that the right CI/CD, source control, and project management processes are in place and that all tools that need to consume data are pointed to the right location.

Finally, we'll complete a business validation. This is meant for you to greenlight the work we've done and confirm that it meets both your expectations and solves the critical business needs we outlined at the beginning of the project.

# CASE STUDY

## INDUSTRIAL MANUFACTURER EXTRACTS BIG EFFICIENCIES FOR BIG EQUIPMENT WITH IOT ON SNOWFLAKE

### THE CUSTOMER

A leading manufacturer of mining and earth-moving equipment sought to increase top line revenue through new products and services, including smart-connected equipment and post-purchase proactive maintenance services. To accomplish this, they needed to transform their existing sensor-based analytics platform into a more efficient, centralized IoT data solution.

### THE CHALLENGE

The manufacturer knew they wanted to take advantage of the latest cloud-native technologies. But they needed help choosing those technologies, executing a successful migration from their existing Hadoop solution, and ensuring the new solution could handle the high volume of IoT data transmitted daily from their equipment sensors.

### THE STRATEGY

phData designed and built a new cloud-native solution for IoT, based on Snowflake, as well as Spark, Kafka, and Microsoft Azure — with automated infrastructure provisioning using infrastructure-as-code, CI/CD for automated deployment, and an architecture that supported dynamic scale and fault tolerance. Then they helped the manufacturer successfully migrate their application from Hadoop to Snowflake, validating the new platforms in-production viability.

### THE OUTCOME

The manufacturer transformed what started out as a small web application into a unified IoT data store, analytics, and visualization platform — designed and optimized by phData to maximize the value of Snowflake's cloud-native architecture.

- Production-tested foundation for enterprise IoT, built on Snowflake
- Cloud-native efficiencies and simplified management
- More unified data for improved collaboration
- Simplified security with Azure AD

By the numbers:

- Daily IoT data...
  - 8-10 billion sensor records
  - 2 million alarm and event records
  - 1 million KPI-derived values
- Historical data intake...
  - 40.8 TB
  - 127 tables
  - 3.8 trillion rows

[CLICK HERE TO READ THE FULL CASE STUDY](#)

# GETTING STARTED

Given the complexity of Hadoop migrations, consider seeking help from phData with expertise and years of hands-on experience. phData specializes in data, ML, and long-term success with Snowflake, and is proud to have been named Snowflake's 2020 Emerging Partner of the Year.

## MIGRATION WORKSHOPS

To kick off a migration plan, we recommend a Migration Workshop with both phData and Snowflake. This will bring our experts together with your team to outline the goals of the project, to begin understanding the current systems landscape, to identify critical components and dependencies, and to discuss the appropriate methodologies and processes highlighted in this document.

### Migration Workshop Agenda:

- Customer use case and goals overview
- Identification of migration and customer challenges
- Overview of methodology and process
- Detailed migration case study and architecture

Duration: 90-120 minutes

Audience: Engineers, Architects, Leaders familiar with Hadoop use-cases and future organizational goals.

[CLICK HERE TO REQUEST MORE INFORMATION ABOUT HOW A MIGRATION WORKSHOP COULD BENEFIT YOUR ORGANIZATION](#)