WHITE PAPER

Best Practices for Hadoop: A Guide From SAS Customers

sas
THE POWER TO KNOW.

Contents

Introduction

As Hadoop becomes commonplace, many organizations have come to learn what "the Hadoop skills gap" really means. Often, organizations don't have the resources or skills needed to effectively adopt and manage Hadoop. The market clearly understands the benefits of Hadoop – the struggle now occurs at a tactical level that involves how to go from initial planning to production.

While every organization is different, and their approaches to Hadoop vary, any organization can learn from the trials, tribulations and accomplishments of others. The journey to begin using Hadoop involves technical, procedural and process challenges. But those who plan well can minimize disruptions and realize significant gains.

Many organizations have used SAS® to help ease the transition to Hadoop. This paper highlights best practices identified by SAS customers using Hadoop. It serves as a guide for others looking to make Hadoop part of their organization.

Planning

Tip 1: Know What Result You Want Before You Start

The first step toward success with Hadoop is knowing what objective your organization hopes to achieve by deployment. Below are six common objectives SAS customers have identified, along with related considerations:

- **The desire to use Hadoop to process data as fast as current storage mechanisms and processing techniques allow**. A common proof-of-value (POV) challenge is to identify where more transformative processes need to be applied to both data structures and processing.
- **The need to use Hadoop as warm storage**. This includes migrating data from backup or external systems into a central store. A good example would be moving data from mainframe backups to Hadoop storage. The business driver here is data accessibility.
- **The ability to keep all your data in one Hadoop environment**. Doing this involves moving data from various sources into Hadoop and then using Hadoop as the source for data access.
- **Being able to process against the data stored in Hadoop**. If the data moves into Hadoop then the data processing is expected to move as well. Some users have worked on transforming SAS code into scoring models and DS2 execution for in-database processing.
- **Reduced costs**. Be sure to ask what overall cost savings you're expecting with Hadoop versus your current data storage and access methods. Many users find that bigger savings come from the results of using all the data stored in Hadoop.
- **Desire to get ahead of competitors**. There are disruptive advantages to the analytically useful information you can get out of Hadoop. Consider this benefit as you make your plans.

Tip 2: Identify and Onboard Users

To succeed with Hadoop, it's important to follow appropriate processes to identify members of the users community and give them access to the Hadoop environment. User identification and onboarding processes need to be in place ahead of time. Many organizations have found these practices helpful:

- **Identify the data first**. To identify the user you must also identify the data he or she needs to access in Hadoop. This might include data stored in SAS data sets, relational database management systems (RDBMS) and other data storage locations. Rather than requiring users to provide details, SAS can help you do a programmatic assessment to identify data usage.
- **Create a secure environment**. Once you have identified the data that's needed, you must create a secure environment within the Hadoop ecosystem. There are some concerns about the security of data fields, data at rest and data extracted from Hadoop. It will save time in the long run if you have a detailed plan and a team in charge of implementation before onboarding users or data.

Tip 3: Solidify Support and Engage Active Participants

Many organizations fail with Hadoop due to undersupported or understaffed projects. SAS customers have identified several best practices that have helped them be more effective with these efforts:

- **Executive sponsorship**. You need both executive sponsorship and technical leadership. The technical lead should be at a high enough level in the organization to engage both IT and the business unit as contributors to the decision-making process.
- **Time commitments**. It's important to have time commitments from the users community that has a vested interest in your SAS and Hadoop implementation. These users should understand their data and know how SAS interacts with the data you intend to put into Hadoop.
- **A vision**. Set a vision, develop a plan, assign tasks and create a timeline based on your success criteria. You need to put some structure around your punch list to avoid inevitable scope creep.

The market clearly understands the benefits of Hadoop – the struggle now occurs at a tactical level that involves how to go from initial planning to production.

Assessing and Mapping the Data

Data assessment and mapping might be one of the least funded and most problematic issues for Hadoop. Don't underestimate the importance of developing a comprehensive data strategy before using Hadoop. Below are some questions you should ask, and strategies you can follow, to help with the data identification and ingestion process:

- **Do you plan to load SAS data into Hadoop?** If so, consider the following example:
  - Is a 10 numeric column, 500,000 row SAS data set considered "big" data? For Hadoop, the answer is no. Your SAS data set, which is 40 million rows, would be represented as a single data split in the Hadoop environment. (In other words, single-threaded processing occurs on one piece of data in Hadoop.) Given this fact, a SAS process running against this SAS data set would be the best performer. Note that Hadoop data splits are 128 million rows and up; so your SAS data set in Hadoop should be many multiples of the data split size before you consider Hadoop for data storage and processing.

  - How do you plan on processing the SAS data you've loaded into Hadoop? If the process is read only, evaluate the type of Hive table you have created. This would include column types, Hadoop storage format and access patterns.
- **Are you planning to load data from a DBMS into Hadoop?** If so, and if the DBMS uses a complex data model, consider how Hadoop is going to interact with that model. Otherwise your ability to port and efficiently process using HiveQL might not work; in that case, you might need to convert the data model into one that can be processed in Hadoop. The way you map the needed data should mirror the processes you plan to run on it.
- **Is data cleansing part of your assessment?** If not, it should be. As you load data from external sources into Hadoop, consider adding cleansing operations as part of the process. With Hadoop, you will find it's much easier to cleanse on the way in than try to change data after it's in place.
- **How are you planning to refresh the data you're loading?** Incremental refreshes might be difficult to implement given that Hadoop has yet to become fully ACID (atomicity, consistency, isolation and durability) compliant.
- **How are you planning to access the data in Hadoop?** Hadoop is at its best when data is processed in large chunks, not individual records. If you primarily need to process individual records, Hadoop is probably not the ideal platform to use.
- **What type of storage format do you plan to use for your Hadoop data?** Although it's widely used, text might not be the best option from a performance standpoint. If your organization plans on accessing the same data using components such as Hive and Impala at the same time, Apache ORC might be a more sensible choice. If Impala is your only data access tool, Parquet might provide the best performance.
- **Do you plan to compress your data in Hadoop?** Several compression options are available: Evaluate the pros and cons of each based on your needs before you make your final decision. And remember that some storage formats like ORC already have built-in compression; knowing this might simplify your decision-making process.
- **How are you planning to secure your data?** By default, Hadoop is a nonsecure environment. Keep in mind that too much security can pose performance issues (this tends to be the case with Apache Knox, for example).
- **Have you considered the encryption zone?** Do you plan on creating pockets of data for specific users or divisions within your organization?
- **Do you plan to implement a data archival process** to phase out old data while ingesting new data? Where will the old data be archived? How "old" is old?
- **Have you thought about disaster recovery scenarios?** Can your Hadoop data pool be rebuilt using other data sources? If not, do you have a backup/recovery strategy in place?

Supporting the Infrastructure

With any Hadoop implementation, you'll need to consider how much investment will be needed for hardware, networking bandwidth and software. But SAS customers have found that having access to dedicated experts and administrators is also vital. Make sure you have the following experts in place to support the infrastructure:

- A SAS administrator – one who understands SAS system requirements for Hadoop, SAS metadata, SAS In-Database, SAS/ACCESS® software, performance and tuning.
- A Hadoop administrator – one who understands Hadoop security, SQL, Hadoop cluster performance, tuning and monitoring.
- Network/security expertise – someone who can assist with user security concerns and configurations as a precursor to enabling security in Hadoop. For example, this expert could address Kerberos, interactions between users and Hadoop, security guidance for establishing Hadoop best practices, and help with Kerberos ticket generation or troubleshooting.
- Hardware/operating system expertise – someone who can help with UNIX or Linux issues, options, installation and patches to meet both SAS and Hadoop system requirements.
- A technical project manager – someone who can provide technical leadership for users. This should include securing resources from the above experts, and providing user support. SAS customers have seen better results in cases where the project manager had a deep understanding of the data and processing goals.

Educating and Training Users

SAS customers have employed several different types of user training. Following are the top three:

- **Functional training for experienced SAS programmers whose data is moving to Hadoop**. It's important to provide this training at the right time. Our customers have done best when the training was followed closely by execution against the new environment. In cases where there was a large gap of time between the education process and execution, results were not as good.
- **Best practices for users as part of the education process**. These practices include SQL optimization, SAS execution strategies and coding efficiencies specific to certain user environments. SAS customers have obtained good results by injecting best practices into user executions against SAS and Hadoop environments.
- **Peer-to-peer training during the knowledge transfer process**. In this scenario, a group of power users experiments with implementations in Hadoop. These experiments then result in best practices and/or mentoring for other users in the same department or organization.

> The problem with data migration is not loading data into Hadoop; it's determining how you're going to manage the data after the fact.

Migrating Data

The problem with data migration is not loading data into Hadoop; it's determining how you're going to manage the data after the fact. To overcome these issues, many organizations have developed a level of sophistication around data organization within their Hadoop environments. This includes having processes that help zone, stratify or layer the Hadoop data store to create an environment that assures data access. Migrating data into this environment is just the start of the journey to having accessible, usable data. Table 1 shows examples of how organizations have handled data ingestion issues with Hadoop.

Source Data	Requirement	In Hadoop	Processing Plan
SAS	Migrate to Hadoop.	Store in SAS Scalable Performance Data Engine format, SASHDAT, Hive table and HDFS file (TXT).	Access this data from SAS, which requires all SAS metadata to be stored with the data. This would include formats, informats, labels and so on.
DBMS	Migrate to Hadoop and develop an update strategy to keep migrated data current. If the data requires transformation to be processed in Hadoop, then the procedure must be preserved. Note that transformation of tabular data into a Hadoop consumable form might be required for complex data models.	Recognize that data is to be joined to other tables migrated from the DBMS, and consider storage size in Hadoop, along with resource requirements. Also consider Apache ORC storage type, data partitioning and other Hadoop constructs.	Perform scoring or processing in Hadoop with the potential final processing on a SAS server.
Stream	Capture web log or other raw data in Hadoop. Build a data processing and organization plan to ready the data for analytics. Maintain original data for auditability.	Zone the data so the data of record is preserved in a highly compressed form. As data is migrated to other zones in the cluster, it will be made available for user consumption.	Preprocess data in Hadoop to cleanse, organize and prepare the data for analytics. The transformation processes are recorded and preserved for auditability. Once the data is placed in accessible zones, the users community can process against it.

Table 1. Requirements and processes to consider related to ingesting data into Hadoop.

Converting Programs

Once the data is loaded, processed and organized in the Hadoop environment, it's time to convert SAS programs or processes to use it. The code conversion process could be as simple as using new LIBNAME statements or as complex as a code rewrite for in-database processing. The table below provides guidance during code conversion.

Requirement	Action	Potential Pros
Access the data in Hadoop from my SAS job.	Some data might not have been moved into Hadoop, so code conversion might first require code review. After the analysis process, all SAS code components whose data has been moved to Hadoop must point to Hadoop. This is done by changing SAS LIBNAME statements and potentially PROC SQL code as well.	Time to working with Hadoop data is shortened. Minimal SAS code change. Quick data validation of the data migrated to Hadoop. Quick identification of performance issues.
Develop and execute a scoring model inside Hadoop.	Modify or write SAS code to enable it to execute inside the Hadoop environment.	Performance gain from in-database execution. Reduced network impact outside of the Hadoop cluster. Reduced SAS storage required for job execution. Ability to score significantly larger sets of data without data extraction.
Run SAS procedures in Hadoop.	Enable SAS procedures – options sqlgeneration=dbms;	Enables PROC FREQ, PROC REPORT, PROC SORT, PROC SUMMARY, PROC MEANS, PROC TABULATE and PROC TRANSPOSE to run advanced HiveQL in Hadoop. Improves performance in listed base procedures.

Table 2. Program conversion considerations.

Proving the Value

SAS has worked on POVs that have specific planned activities, as well as some that are more dynamic. Consider the following ways to add value to your POV:

- Many organizations have dynamic POVs where the use cases have not been fully developed. They know they want to process data in Hadoop, but are not fully aware of what they want to do with the data in Hadoop afterward. Dealing with a mass of data and time constraints is also problematic. If this type of POV is required, consider specific SAS processes against generated data or identified user data as a case study. The dynamic POV can then shift away from specific user processing to POV processes. This POV could deliver on requirements on tight time schedules.

- Many SAS customers have planned POVs where data, programs, events and checkpoints have been defined. But even with careful planning, the data loading, data organization in Hadoop and tests have caused delays. Experience with planned POVs proves that it's best to test early. For example, once the data is loaded, the interaction with the data can be tested ahead of user programs. User programs can be reviewed and problems can be identified before running against Hadoop. The plan and timeline for many companies is very rigid, but the organized support and the POV for Hadoop needs to be flexible.

So, what is the best approach to complete the POV and move on? Many SAS customers have found that investing technical resources in the design and process are critical. If you're considering moving straight from a POV to production, put the investment in the POV so you can develop processes that can be applied to production. We have observed that those who rush from POV to production often run into issues and time delays that should have been identified in the POV.

Moving to Production

A production environment can be the finish line or the start of the race. Many organizations struggle with production environments because they've missed some fundamental concepts. Below are some best practices that SAS customers have found helpful when they were ready to move to production:

- **Consider the size of the production environment**. For a multitenant Hadoop and SAS environment, how was the sizing done? What is the impact of the users on the Hadoop environment? Who is going to help identify and resolve these issues?
- **Are the user onboarding and security processes complete?** As different user groups transfer to the production system, the onboarding process must be production quality. The impact of the data and the processing requirements of new groups might require production upgrades.
- **Have you planned for data updates?** Specifically, how will data in a production environment be created and maintained? Do you have a plan for users' data ingestion needs and requirements?
- **What is your disaster recovery plan?** How is your data maintained at both on- and off-site locations, and what are your disaster recovery processes? Do you have service-level agreements (SLAs) for system uptime, and how does Hadoop play into those scenarios?
- **Have you planned for data security?** Do you have procedural or process requirements for data at rest, data on the wire or duplicated data via Hadoop extraction?
- **Have you determined how to measure user satisfaction?** The experience the users community has in a production environment needs to be great. A clean migration from the POV environment – that is, a smooth onboarding of data, users and processes to a production environment – is critical.

Post-Implementation: Putting Strategies Into Practice

Enabling Data for Use

After establishing a plan for identifying, collecting, cleansing, using and loading data into Hadoop, it's time to execute. Following are some of the top considerations at this stage:

- The data is in and it's time to run validation and performance tests against the data that's going to be consumed by SAS jobs and processes. You must have assessments of run times and appropriate actions ahead of the user migration.
- How will you manage and process data created in Hadoop by SAS jobs? An example would be a process running against a Hadoop table that will produce another Hadoop table. Is data creation permitted? Where will the data go and how will it be consumed? How will it be shared and how is it managed? Best practices for cleanup must be put in place – not only to save space but to also keep SAS jobs and processes from failing.
- Do you know how you will keep data current in the Hadoop environment? You should consider this before you load the first record. Many terabytes can get loaded multiple times due largely to lack of understanding the data and planning. Full ACID compliance requires you to deploy innovative processes that keep your Hadoop data current.

Those who rush from POV to production often run into issues and time delays that should have been identified in the POV.

Securing the Environment

The topic of security gets resolved at the last minute for many. Planning ahead for security is, of course, a better approach. With Cloudera and Hortonworks, these security scenarios are often used:

- Hortonworks – Kerberos LDAP(S) and Ranger.
- Cloudera – Kerberos LDAP(S) and Sentry.

Many organizations run Hadoop environments inside their data center, while others run Hadoop in the cloud. Some organizations may have a mix of both. Managing security in these different environments can be challenging. What has worked for most organizations is to have in-house or vendor expertise with Kerberos and the related Hadoop security structure. Some planning and testing of the infrastructure – including SAS and Hadoop interactions – have helped develop the plan for user onboarding.

For example, SAS, relies on operating system components to generate Kerberos tickets that allow you to access your Hadoop environment. The infrastructure in your environment must support the ticket generation process. Valid tickets are required from all users who need to access your Kerberized Hadoop environment.

Once you've gained access to Hadoop, data access can be controlled using Ranger or Sentry. To save time, consider these options and decide how you want to enable users and secure the data in Hadoop before you start implementing security.

Dealing With Users Who Are Resistant to Change

It's vital to minimize internal disruptions as you begin using your SAS and Hadoop environment so you can get more value from your data. You should understand and develop some fundamental concepts and practices before engaging with users. There's no magic in doing this – but a shared commitment will help you achieve success in this data environment. The perception you instill in your users and consumers from the first step will help you deliver ROI.

Following are some suggestions that may help ease the user transformation process as you move to the new SAS and Hadoop environment. Table 3 shows an action that was taken, the user involvement, user responses to the action and the overall outcome. As you will notice, user input is vital.

Action	User Involvement	User Responses	Outcomes
Unload data from various sources into Hadoop, then use this data in your SAS jobs.	Users have minimal direct interaction. Access to some documentation and some knowledge transfer is provided. SAS administrator is asked to help.	General resistance. Lack of time. Inability to identify and resolve performance issues. Failure (that's communicated to others).	Failure to achieve the goal. Difficulty regaining trust and traction.
Gather data to load Hadoop for SAS uses by querying the users community for data and process requirements.	Meetings are held to let users know what will be happening. Requirements gathered. Knowledge transfer takes place. SAS administrator is asked to help.	Users provide the bare minimum of information, at the last minute. Context is missing for cross-functional requirements in the same department. Implementation delays occur due to inconsistent or erroneous data.	Disappointing degree of wins for time invested. Big data strategy remains suspect and unproven. User commitment wanes.
Perform data usage analysis for specific user groups. Identify data use scenarios that are a good fit for the Hadoop environment. Prioritize user groups that are most affected and ensure that their goals mesh with corporate needs/priorities. Share information about data analysis rather than asking about wants or needs.	Steering committee forms, made up of decision makers, IT and select users. Data transformation, user process identification, project timeline and success criteria are established. A technical team of SAS, Hadoop and system administration staff answers questions and resolves problems. Before rolling out to users, it's proven that the system can handle the load. Before starting, training, knowledge transfer, best practices and procedures are provided. A feedback process is in place.	Resistance, but with a clear understanding of why, when and how learning curve slope increases. A feeling of less stress in trying to get things to work as support processes are specified. Increased interest as users realize performance gains with new SAS coding and execution techniques.	Some wins are identified and achieved. Procedure and process improvements are noted. Usage scenarios and best practices become clear. You're ready to queue up the next set of users/big data usage scenarios based on corporate priorities.

Table 3. A user resistance matrix for adoption of a SAS and Hadoop environment.

Entering POV/Production Phase in a Multitenant Hadoop Environment

Once the POV is complete, it's time to move forward and prove the concepts in a production environment. The production concept can include both SAS and Hadoop components that are shared across your organization. Some organizations have taken the next steps quickly. But in hindsight, they say it would have been better to follow these steps:

1. Identify the overall load that will be placed on your production environments. This can include number of jobs, interactions, data extractions, data loads, and number of users in both the ad hoc and batch categories.
2. Identify any priority jobs that must execute within a specific SLA. This would be the jobs that you are adding to the environment. It's a best practice to identify those jobs with requirements before migration.
3. Identify data, processes and functionality that must be considered as part of your disaster recovery strategy.
4. If possible, test a simulated load on the production environment. This would give your administration staff a heads up about the needs of the new users community.
5. Identify Hadoop queues, resource management, capacity management, user onboarding and the security impact before you move to the new environment.
6. Upgrade infrastructure in a test environment, implement performance tests and – upon success – move to a production environment.

You might not have the infrastructure or bandwidth to validate the full impact of moving to production. But knowing that the impact is coming can help enhance the monitoring process.

Conclusion

This paper is just a starting point for effectively moving to and using SAS with Hadoop. Putting these best practices in place requires a commitment of time and resources. The first step should be to analyze what you are trying to achieve by making this change.

Data is one of the cornerstones of success with Hadoop – and it's essential to know how that data will be processed in Hadoop. Organizations have found that getting the appropriate data transformed and loaded into Hadoop ahead of user adoption is a best practice.

Once the data is loaded, you'll want the first user experience to be transparent and require minimal change. The more disruptive changes to coding and processes will be easier to introduce through knowledge transfer – and this should be based on results of your independent testing in Hadoop. As users become more engaged, proof from actual results will support the rationale for doing things the new and better way.

Your journey through the different implementation milestones of your Hadoop project should provide rich knowledge for your organization. The first group of Hadoop users will identify various problems, issues, concerns and questions. By the time the second group begins, they should not be providing the same feedback on the same issues. This continuous learning process is what will lead to Hadoop success – and ultimately success for your company as a whole.

Learn More

Every organization is unique, and approaches to Hadoop will always vary. But the best practices our customers have identified can provide an excellent road map for others who are just beginning their Hadoop journey. Learn more about SAS and Hadoop by visiting sas.com/hadoopsolutions.

To contact your local SAS office, please visit: sas.com/offices