Paper SAS378-2017

How SAS® Customers Are Using Hadoop – Year in Review

Howard Plemmons Jr., SAS Institute Inc.

ABSTRACT

Another year implementing, validating, securing, optimizing, migrating, and adopting the Hadoop platform. What have been the top 10 accomplishments with Hadoop seen over the last year? We also review issues, concerns, and resolutions from the past year as well. We discuss where implementations are and some best practices for moving forward with Hadoop and SAS® releases.

INTRODUCTION

The ability to learn from the trials, tribulations, and accomplishments of others is a competitive business advantage. Taking some of the insights provided in the paper and applying them to your SAS Hadoop implementation or planning efforts can help you achieve results at a quicker pace. The paper contents apply to both SAS Hadoop use under consideration, design, implementation, or Proof of Value (POV) implementations.

The journey of using SAS with data stored in Hadoop has technical, procedural, and process challenges; however, for those that plan well the disruptions can be minimized and gain realized. This paper goes beyond introducing the SAS Hadoop concept and is intended for those that are in design or implementation.

The top 10 accomplishments mentioned in the abstract have morphed into ten focus areas and then some implementation examples from our work with SAS customers. The focus areas seem straight forward; however, the repetitive process that some use makes them more difficult than they need to be. An example would be moving quickly to an environment without answers to basic questions, identified goals, or expected results. The repetitive process is trying to make what you have locked-in on meet the expectations that are not yet set.

WHAT ITEMS OR EVENTS MADE THE TOP TEN

The formulation of a top 10 list was based on SAS customer experiences with Hadoop. The items chosen concentrate on issues uncovered during discovery, design, POV, and production implementations of SAS and Hadoop. The list below is not comprehensive or in any specific order:

WHAT IS THE DESIRED RESULT?

What is the strategy for deploying and using a SAS Hadoop environment? Here are a few we have worked on with SAS customers this year:

- Prove that SAS Hadoop can process as fast as current data storage mechanisms and processing techniques. A common POV challenge where more transformative processes need to be applied to both data structures and processing challenges.
- Hadoop as warm storage. This would include migrating data from backup or external systems into a central store. A good example from several users would be moving data from mainframe backups to Hadoop storage. The business driver here is data accessibility.
- Keeping all the data in one Hadoop environment. This would include moving data from various sources into Hadoop and using Hadoop as the source for data access.
- Process against the data stored in Hadoop. If the data moves into Hadoop then the processing on that data is expected to move as well. Users have worked on transforming SAS code into scoring models and DS2 execution for in-database processing.

- Cost savings. Some have started the Hadoop expecting overall cost savings over current data storage and access methods. Many have found that true savings is obtained through the results obtained using data stored in Hadoop.
- Realizing the disruptive benefit of being able to get analytically ahead of the competition.

SOLIDIFY SUPPORT AND ENGAGE ACTIVE PARTICIPANTS

We have seen that under supported or under staffed Hadoop projects have a high probability of delay or failure. Here are some team concepts that have been effective:

1. You need executive sponsorship and you need to have technical sponsorship/leadership. The technical lead should be at a level in the organization to engage with either IT or the business unit as contributors to the decision making capacity.
2. You need time commitments from the end-user community who have a vested interest in SAS Hadoop implementation. These users should understand their data and have a good understanding of how SAS interacts with the data you intend to put into Hadoop.
3. Set the vision, develop the plan, assign tasks, and create a timeline based on success criteria. You need to put some structure around your punch list to avoid inevitable scope creep.

DATA ASSESSMENT AND MAPPING

Data assessment and mapping might be one of the least funded and most problematic issues with Hadoop. The investment of time to develop a comprehensive data strategy when using Hadoop is critical. Some of these questions and strategies have helped in the data identification and ingestion process:

1. Are you planning on loading SAS data into Hadoop? If so, consider the following example:
    a) Is a 10 numeric column, 500K row SAS data set big data? For Hadoop, the answer is no. Your SAS data set, which is 40M in size will be represented as a single data split in the Hadoop environment. What this means is single threaded processing on one piece of data in Hadoop. Given this fact, your SAS process running against this SAS data set is the best performer. Note that Hadoop data splits are 128M and up; therefore, your SAS data set in Hadoop should be many multiples of the data split size before considering Hadoop for data storage and processing.
    b) How do you plan on processing the SAS data you have 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.
2. Are you planning on loading 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. Without consideration, your ability to port and efficiently process using HiveQL might not work. If that is the case, then converting the data model into one that can be processed in Hadoop might be necessary. How you map the needed data should mirror the processes you plan to run on it.
3. Data cleansing should be part of your assessment. As you load data from external data sources into Hadoop consider adding cleansing operations as part of the process. With Hadoop, you will find it is much easier to cleanse on the way in rather than trying to change data in place.
4. How are you planning to refresh the data you are loading? Incremental refreshes might be difficult to implement given that Hadoop has yet to become fully ACID complaint.
5. How are you planning to access the data in Hadoop? Hadoop is at his best when data is processed in large chunks as opposed to individual records. With the latter scenario, Hadoop might simply prove not to be the ideal platform for your organization.

6. What type of storage format are you planning on using for your Hadoop data? While 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, ORC might be a more sensible choice. If Impala is your only data access tool, Parquet might provide the best performance.
7. Are you planning on compressing your data on Hadoop? Currently several compression options are available. Evaluate pros and cons of each one based on your needs before making a final decision. Also remember that some storage formats like ORC already have compression built-in, which might ease the decision-making process.
8. How are you planning to secure your data? By default, Hadoop is a non-secure environment, but on the other hand, too much security can pose performance issues (KNOX, for example).
9. What about encryption zone? Do you plan on creating pockets of data to specific users or divisions within your organization?
10. Do you plan on implementing a data archival process to phase out old data while ingesting new one? Where will the old data be archived? How "old" is old?
11. Have you thought on how to handle 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?

INFRASTRUCTURE INVESTMENT

We see two parallel tracks on investing and tracking infrastructure investment. The first is hardware, networking bandwidth, software, and adherence to your internal protocols that govern these areas. The second is the expertise that is needed to implement and maintain a SAS Hadoop environment. Dedicated access to these experts and administrators can have a positive impact on your implementation timeline. This administration expertise includes:

1. SAS Administrator – one who understands SAS system requirements for Hadoop, SAS metadata, SAS in-database, SAS/ACCESS, performance, and tuning.
2. Hadoop Administrator – one who understands Hadoop security, SQL, Hadoop cluster performance, tuning and monitoring.
3. Network/Security Expertise – one who can assist in user security concerns and configurations as a precursor to enabling security in Hadoop. Some of the items would include Kerberos, interaction between users and Hadoop, security guidance for establishing Hadoop best practices, help with Kerberos ticket generation or troubleshooting.
4. Hardware/OS Expertise – one who can help with UNIX or Linux issues, options, installation, and OS patches to meet both SAS and Hadoop system requirements.
5. Technical Project Manager (PM) – one who can provide end-user technical leadership, which would include securing resources 1-4 and end-user support. We have seen increased success when the PM has a greater understanding of the data and processing goals.

END-USER IDENTIFICATION AND ONBOARDING

To be successful the procedure and process of both identifying members of the user community and providing them access to the Hadoop environment is critical. Identification and onboarding processes need to be in place ahead of time. You might find some of these practices helpful:

1. Identification of the end user also means identifying the data that they will need in Hadoop. This data might contain data stored in SAS data sets, RDBMS, and other data storage locations. We have helped with programmatic assessments that identify data usage rather than requiring end users to provide the data details.

2. Once you have data identified, you must create a secure environment within the Hadoop ecosystem. We have seen concerns with security with data fields, data at rest and extracted data from the Hadoop environment. Having a detailed plan and implementers before onboarding end users or end-user data will save time.

END-USER EDUCATION AND TRAINING

We have seen several scenarios used for end-user training. Some of these have had different levels of return for their Hadoop implementations:

1. Functional training for experienced SAS programmers whose data is moving to Hadoop. Deciding when to provide training and best practices for the user community should be on your timeline. We have seen good success when the training is followed with execution against the new environment. Training with a large gap between the education process and execution provided diminished returns in some cases.
2. Best practices developed for end users as part of the education process was shown to be very effective. These practices include SQL optimization, SAS execution strategies, and coding efficiencies specific to end-user environments. We have seen good results injecting best practices into end-user executions against SAS Hadoop environments.
3. Peer-to-peer training has been proven effective in the knowledge transfer (KT) process. In this scenario, a group of power users experiments with implementations in Hadoop. These experiments would result in best practices and/or mentoring for other end-users within the same department or organization.

DATA MIGRATION

This critical step has been problematic for many in 2016. We have seen unsuccessful first attempts to ingest data into Hadoop. It is not that you can't load data into Hadoop; it is what you are going to do with it after the fact. To overcome these data issues, we have seen many develop some sophistication around data organization within the Hadoop environment. This includes processes that help zone, stratify, or layer the Hadoop data store to create the environment that can assure data access. Migrating data into this environment is just a start of the data journey with the ultimate goal of usable data. For examples see Table 1 below:

Source Data	Requirement	In Hadoop	Processing Plan
SAS	Migrate to Hadoop	Store in SAS SPD 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 update strategy to keep migrated data current. If the data requires transformation in order to be processed in Hadoop, then the procedure must be preserved. Note, transformation of tabular data into a Hadoop consumable form might be required for complex data models.	Data is to be joined to other tables migrated from the DBMS and storage size in Hadoop is a consideration as is resource requirements. Consider ORC storage type, data partitioning, and other Hadoop constructs.	Scoring or processing is to be performed in Hadoop. Potential result set extraction for final processing on a SAS server.
Stream	Capture weblog or other raw data in Hadoop. Build a data processing and organization plan to ready the data for analytics. Must maintain original data for auditability.	Zone the data so that 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.	Data is pre-processed 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 user community can process against it.

Table 1. Considering Requirements and Processes to Ingest Data into Hadoop

PROGRAM CONVERSION

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

Requirement	Action	Pros	Potential Cons

Access the data in Hadoop from my SAS job	Not all data might have been moved into Hadoop, so code conversion might require code review first. 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 shortened Minimal SAS code change initial impact Quick data validation of the data migrated to Hadoop Quick identification of performance issues	Performance – accessing Hadoop data from SAS jobs might run slower Impact on the Hadoop cluster and network from additional I/O requirements Not using Hadoop in the most efficient way
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	SAS PROCs will not run inside Hadoop Time needed to develop and test the SAS scoring process Requires some training or experience with DS2 Scoring model needs to keep the scoring output inside Hadoop for optimal performance Scoring model management
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	Limitation in procedure options that are supported in the specified mode of SQL generation

Table 2. Program Conversion Considerations

PROOF OF VALUE COMPLETE

The planned POV (Proof of Value) is complete. We have worked on POVs that have specific planned activities as well as on ones that are more dynamic. Consider the following items to add value to your POV:

1. We have dealt with dynamic POVs where the user cases were difficult to come by. Dealing with a mass of data and time constraints for POV is problematic. If this type of POV is required, consider specific SAS processes against generated data or identified end-user data as a case study. The dynamic POV can then shift away from specific end-user processing to POV processes. This POV could deliver on requirements on tight time schedules.
2. We have dealt with planned POVs when data, programs, events, and checkpoints have been defined. Even with careful planning the data loading, data organization in Hadoop and tests have caused delays. The lessons learned from planned POVs is to test early as an assurance for success. For example, once the data is loaded the interaction with the data can be tested ahead of end-user programs. End-user programs can be reviewed and problems identified before running against Hadoop. The plan and timeline have been developed to be static; however, the organized support around the POV for SAS Hadoop needs to be dynamic.

So, what is the best approach to complete the POV and move on? Simply put, investment of technical resources in design and process. If you are considering moving straight from a POV to Production, put the investment in the POV to help develop processes that can be applied to Production. We have seen those who rush from POV to Production run into issues and time delays that should have been identified in the POV.

PRODUCTION ENVIRONMENT

A production environment can be the finish line or the start of the race. It is the culmination of the activities required by your organization for a production environment. We have seen the struggle with production environments that have missed some of the fundamental constructs. Some of these might help you when you're ready to proceed to production:

1. Size of the production environment? For a multi-tenant Hadoop and SAS environment how was the sizing done? What is the impact of the end users on the Hadoop environment? Who is going to help identify and resolve these issues?
2. User onboarding and security process complete? As different user groups transfer to the production system the onboarding process must be production quality. The impact of the data and processing requirements of new groups might require production upgrades.
3. Data updates? Specifically, how will data in a production environment be created and maintained? Do you have a plan for end-user data ingestion needs and requirements?
4. Disaster recovery? How is your data maintained in both on and off site locations and what are your disaster recovery processes? Do you have SLAs for system up time and how does Hadoop play into those scenarios?
5. Data Security? Do you have procedural or process requirements for data at rest, data on the wire or duplicated data via Hadoop extraction?
6. End-user satisfaction? The experience the end-user community has in a production environment needs to be a great one. A clean migration from the POV environment, or a smooth onboarding of data, users, and processes to a Production environment is critical.

PUTTING STRATEGIES INTO PRACTICE

Here are a few examples of SAS Hadoop projects in 2016 that we have helped customers work through. We hope that this list will provide some insight into areas that can disrupt your required delivery of Hadoop ROI.

ENABLING DATA FOR USE

After establishing a plan for identifying, collecting, cleansing, using and loading data into Hadoop it is time to execute. Once you have loaded the data in Hadoop, what's next:

1. The data is in and it is time to run validation and performance tests against the data that is to be consumed by SAS jobs and processes. Assessment of run times and appropriate actions ahead of the end-user migration has become a best practice.
2. 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, how will it be consumed, how will it be shared and how is it managed. Note, if you change a LIBNAME statement in a SAS job to a specific Hadoop user and a specific schema, you are creating Hadoop Hive tables as a result in those hive-schemas. Best practices for cleanup must be put in place to not only save space but keep SAS jobs and processes from failing.
3. Concern with keeping data current in the Hadoop environment must be considered before you load the first record. We have seen many terabytes (TB) loaded multiple times due largely to lack of understanding the data and planning. Full ACID (Atomicity, Consistency, Isolation, and Durability) compliance is a concern that will require that you deploy innovative processes to keep your Hadoop data current.

HOW CAN I SECURE THE ENVIRONMENT?

A hot topic that seems to get resolved at the last minute. We have seen a simple recipe that has worked for many in 2016. Considering Cloudera and Hortonworks we have seen these security scenarios in practice:

For Hortonworks – Kerberos LDAP(S) and Ranger

For Cloudera – Kerberos LDAP(S) and Sentry

We have worked with customers that run Hadoop environments inside and others that run Hadoop in the cloud. How to operate with the different security methods can be challenging but not impossible. What is required is in-house or vendor expertise with Kerberos and the related Hadoop security structure. Some planning and testing of the infrastructure, which would include SAS Hadoop interaction, can help you develop the plan for end-user onboarding.

For example, SAS, relies on OS components to generate Kerberos tickets that will allow you to access your Hadoop environment. The ticket generation process requires that infrastructure is in place within your environment to support it. Valid tickets will be required from all users that need to access your Kerberized Hadoop environment.

Once you have gained access to Hadoop, your access to data can be controlled using Ranger or Sentry. These are tools that give you options for security; however, to save time, decide how you want to enable users and secure the data in Hadoop before you start security implementation.

END USERS ARE RESISTANT TO CHANGE

How to minimize internal disruption while gaining strategic value of your data and the SAS Hadoop environment is critical. Fundamental concepts and practices must be understood and developed before engaging with selected end users. Ideas from Table 3 may help you plan your end-user transformation process to the new SAS Hadoop environment:

Strategy	End-User Interaction	End-User Reaction	Outcome
Unload data from various sources into Hadoop then use this data in your SAS jobs	Minimal direct interaction, access to some documentation and some knowledge transfer (KT), SAS Administrator now tasked to help with Hadoop	Resistance, no time, unable to identify and resolve performance issues, failure and communicated failure to other end users	Failure, hard to regain trust and traction
Query the end-user community for data and process requirements. Use this information to gather data to load Hadoop for SAS use.	Meetings, requirements gathering, KT, SAS Administrator to help with implementation	Provide only info asked for at the last minute, context missing for cross functional requirements in the same department, implementation delays due to inconsistent or erroneous data	Unrealized success for time investment, big data strategy remains suspect and unproven, end-user commitment wanes
Perform data usage analysis on specific end-user groups (for example, what jobs, where is the data, how big is the data, and so on). Identity data use scenarios that are a good fit for the Hadoop Environment (for example, 50 row SAS data set is not, 2TB SAS data set might be). Prioritize user group interaction to corporate needs and priorities. Have information to share about data analysis rather than asking what do you have or need.	Form a steering committee made up of decision makers, IT, and selected end users. Plan the data transformation, end-user process identification, project timeline, and success criteria. Form a technical team of SAS, Hadoop, and System Administration staff to answer questions and resolve problems as needed. Prove the system can handle the load before rolling out to end users. Provide training, KT, best practices, and procedures before starting. Set up a feedback process.	Resistance; however, with a clear understanding of why, when and how learning curve slope increases. Identified and focused support processes lessen stress on trying to get things to work. Realization of performance with new SAS coding and execution techniques builds interest.	Some identifiable and achieved success. Procedure and process improvements identified. Usage scenarios and best practices identified. Ready to queue up the next set of users /Big Data Usage scenarios based on corporate priority/need.

Table 3. Resistance Matrix to SAS Hadoop Adoption

We have seen no magic other than the shared commitment that will help you achieve success in this data environment. The perception you instill in your user base and consumers from the first step will help you deliver ROI.

ENTERING POV/PRODUCTION PHASE IN A MULTI-TENANT HADOOP ENVIRONMENT

The POV is complete and now it is time to move forward and prove the concepts in a production environment. The production concept can include both the SAS and Hadoop components that are shared

across your organization. We have seen the next step taken quickly. In hindsight, 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. Identification of those jobs with requirements before migration is a best practice.
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 to needs of the new user community.
5. Identify Hadoop queues, resource management, capacity management, user onboarding, and security impact before the move to the new environment
6. Obtain and implement infrastructure upgrades before moving production work to the production environment or sensitize the other Tenants of the cluster to performance impacts

You might not have the infrastructure or bandwidth to validate; however, knowing that impact is coming can help improve or heighten the monitoring process.

CONCLUSION

The information provided is just a starting point for effectively moving to and using SAS and Hadoop. To put the mentioned practices in place requires time and resource commitment. You first must analyze what you are trying to achieve before doing so. As mentioned the key to Hadoop success is DATA and how that data will be processed in Hadoop. You will find that getting appropriate data transformed and loaded into Hadoop ahead of user adoption is a best practice.

Once the data is loaded, you want the first user experience to be transparent and require minimal change. The more disruptive changes to code and processes will be easier to introduce through knowledge transfer (KT). The KT is based on results from your independent testing in Hadoop. As the end users are engaged the new and better way is backed up with processing facts.

Your journey through the different implementation milestones of your Hadoop plan should result in KT development for your organization. For example, the first group of users that consume the Hadoop environment will identify problems, issues, concerns, and questions. The second group through should not be providing the same feedback on issues identified earlier.

ACKNOWLEDGMENTS

I would like to recognize the editors and contributors to this paper. This includes the following members of SAS Professional Services and the Database Technology Practice:

Mauro Cazzari

Lee Herman

Rajasree Gutha

Tyler Wendell

RECOMMENDED READING

The following references can be found on the SAS Technical Support website at http://support.sas.com/en/support-home.html :

- *SAS® ACCESS to Hadoop*

- *SAS® 9.4 In-Database Products: Administrator's Guide (9.4 Maintenance <your release of SAS>)*
- *SAS® 9.4 In-Database Products: User's Guide (9.4 Maintenance <your release of SAS>3)*
- *SAS® and Hadoop Technology: Deployment Scenarios*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Howard Plemmons
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
919-531-7779
Howard.Plemmons@sas.com