BI Experts' Perspective

When It's Time to Hadoop

Rob Armstrong, Scott Barnes, Joey D'Antoni, Tracy Ring, and Srinivas Varanasi

Rob Armstrong is a data and analytics enthusiast serving in United Data Architecture Marketing for Teradata.
Rob.armstrong@teradata.com

Scott Barnes is director of information management consulting for Deloitte Consulting LLP.
scbarnes@deloitte.com

Joey D'Antoni is the principal consultant at Denny Cherry and Associates Consulting.
Joey@dcac.co

Tracy Ring is a specialist leader of global alliances for Deloitte Consulting LLP.
tring@deloitte.com

Srinivas Varanasi is managing principal consultant at Excelion Consulting.
svaranasi@excelionconsulting.com

Q **Meagan Hayes is the BI director at Southern Style, a leader in upscale clothing for men and women. It has two manufacturing plants and sells online and through an international network of stores. Meagan leads a team of six BI professionals who manage the data warehouse, develop enterprisewide applications, and help various business units create specialized applications.**

To date, most of the BI team's work has been "vanilla," that is to say, traditional data warehousing technology, data access and analysis tools, and applications (e.g., reporting, dashboards, and scorecards). Meagan and senior management recognize that more can and should be done in terms of the storage and analysis of new big data, including social media and Web log data.

Meagan senses that there are opportunities to use Hadoop as a cost-effective way to ingest, store, and analyze data. She does have questions:

- **Does it make sense to store all data that might potentially be used in a Hadoop-based data lake, including the old and new data sources?**
- **Meagan has well-developed ETL processes for the data warehouse. Should she use Hadoop for all ETL processes?**
- **Can the set of BI tools currently used (e.g., Tableau, MicroStrategy) to access warehouse data connect directly to Hadoop?**
- **Southern Style provides mobile BI to executives, regional managers, and store managers. Does Hadoop fit into mobility?**
- **Meagan is also thinking about sandboxes. Does it make sense to use Hadoop as a sandbox?**

ROB ARMSTRONG

It is when to use which, not which one to use.

Meagan should be looking at the options available to her because much has changed from the days of SQL analytics of transactional data. Although it is true that more can be done with "big data," Meagen should realize that more can be done with her current data beyond the "vanilla" work done to date. There are plenty of advanced analytics and processes she can leverage from her data warehouse. However, where a relational database is not the best answer, options must be considered.

This is the critical point. Hadoop is attractive because of its scale and parallelism, seemingly lower cost, and especially its ability to work with nonrelational data. Many companies focus on the apparent low cost because the software is free, but there are hidden costs such as hardware and software support, development efforts, and daily management to name a few. Meagan's focus should be on a use case for leveraging nonrelational data where using a relational database is not the right choice.

A relational database provides an environment with all the ACID properties for data management, a query optimizer, workload management, high concurrency, and the ability to effectively and easily manage integrated data models. Meagan should not dismiss the good work done in her data warehouse but should complement that work with the analysis of multistructured data in Hadoop.

Without knowing details of the database or Hadoop vendor, I suggest that Meagan work closely with all her vendors (including BI and ETL tools) to get the best value from each tool.

Data Lake and ETL

The real value of the data lake is quick ingestion and the storage of nonrelational data or data that you do not want to take the time and effort to transform and model into the data warehouse. However, much of the data coming in is already well known and relational, so not all ETL processes need or should be changed to Hadoop.

If the current processes run well and the data will ultimately be loaded into the warehouse, the cost to recode it is clearly not worth the benefit. A small percentage of jobs may be in need of upgrading or revisiting the process, and at those times, Hadoop should be considered.

With other data, the economics of data management have changed enough that Meagan can now store much of her "dark data" at a low cost within the data lake; that data should be stored in its raw form.

The important part, though, is that the data lake still needs some level of governance including security, privacy, and metadata. Otherwise it will become a worthless data swamp. Without that proper construction and oversight, Meagan will end up with the problem of data everywhere and with little understanding of its value to the business.

BI Tools and Mobility

Part of the low cost of Hadoop is the "schema-on-read" philosophy, where data is not modeled (or is only lightly modeled) on ingestion. The more you want to read the data, the higher the cost if you have schema-on-read.

That is the trade-off here because the tools are generating SQL and expecting structure and an optimizer to maximize performance. SQL on Hadoop is getting better but queries of complexity may be beyond the scope today. This means that tools either need to pull much more data back for local processing or the Hadoop data needs to be preprocessed after the ETL jobs to provide the schema that BI can leverage.

Mobility is a closely related topic because mobile devices are just running a different type of BI. Meagan can take the mobile requirement and create answer sets that can reside in memory for fast response.

Can all this be done? Yes, but how much work does Meagan want to spend getting the data properly aligned in Hadoop to make it happen? The more she wants Hadoop to provide database capabilities, the more the data warehouse is the answer.

Sandboxes

Sandboxes are good for giving users the ability to explore and test ideas without IT involvement. Either the data warehouse or Hadoop environment can provide that ability. The question is the same as before: What type of data are users working with and what is their skill set?

If the data is already modeled and relational and users know SQL, then use the database. If the data is going to be loaded without modeling or is nonrelational and users are adept at programming, then Hadoop is a better option.

The bottom line: any technology used incorrectly will be inefficient and expensive, so it is best to understand the value each brings and leverage it in a well-planned ecosystem of data storage and data analytics. This is how Meagen will provide the best analytical capability at the lowest infrastructure cost to the business community.

A SCOTT BARNES AND TRACY RING

Although Meagan and her senior management team at Southern Style are asking the right questions with respect to Hadoop and big data, they need to expand their thinking beyond the the storage and analysis of "new big data."

A good place to start is to look at all the potential use cases and opportunities for this evolving technology. Hadoop is not just a low-cost replacement to existing technologies but an incremental addition to traditional architectures.

> The bottom line: any technology used incorrectly will be inefficient and expensive.

In the next-generation analytics architecture, we typically see valuable data mined out of the Hadoop environment and pushed to an analytics environment where the tools are optimized to connect to (and perform from) a query perspective.

Although there are new technologies (such as Apache Spark) that allow analytical environments to be built within the Hadoop environment, this space is still evolving and probably too "bleeding edge" for a company like Southern Style.

When we think about how Southern Style could leverage Hadoop, we see four major categories of use cases.

Use Case #1: EDW Augmentation

Example: Clickstream analysis that could provide the "next best offer" for an online channel.

Characteristics: Hadoop is primarily used to perform ELT (vs. ETL) functions and large volumes of unstructured/semistructured data are standardized for downstream consumption. Information may be accessed directly or extracted from the new data store and merged with existing traditional data sets for analysis.

Use Case #2: Integrated Analytics Platform

Example: Customer segmentation or consumer sentiment analysis using social media feeds.

Characteristics: In this use case, Hadoop is integrated with traditional toolsets to drive deeper analytics using unstructured and structured data together. Result sets are made available for further reporting and analysis directly from Hadoop and/or passed on to traditional tools.

Use Case #3: Real-Time Decision Platform

Example: A real-time platform providing live customer recommendations and personalized offers to Southern Style shoppers while they're in or near the store.

Characteristics: Here, Hadoop is used to analyze events in real time and analytics are embedded into operational processes to support frontline decision making.

Use Case #4: Warm-Cold Data Store

Example: Archiving Southern Style's historical point-of-sale transactions.

Characteristics: Hadoop is used to archive high-volume/low-value data. Formatted information is moved from the data warehouse to Hadoop, and data is made available for use in analytics sandboxes or through reporting tools directly from Hadoop.

With respect to Meagan's question about Southern Style storing all of its data in the lake, the answer is "no." Typically, there is no need to move structured data that is already available to the end-user community via the data warehouse back to the data lake. The primary purpose of the lake is to acquire and store new data sources and provide an environment to explore this data, with the goal of identifying relevant patterns and insights in the data and moving that "forward" in the architecture to an area optimized for query and analysis.

In reference to Meagan's ETL processes, the answer is "probably not." In the early years, Pig, Python, and other Hadoop-centric data manipulation tools were required to manipulate data in Hadoop, but these technologies face similar issues to those that led to the rise of traditional ETL platform vendors: scalability, metadata traceability, reusability, and maintainability.

Many of the traditional ETL vendors have developed big data connectors, such as Informatica Big Data Edition, Oracle Big Data Connector, and SQL Server SQOOP Connector, which leverage ETL's underlying technology and associated benefits while integrating with Hadoop, thus providing the best of both worlds.

> The cloud overcomes many of the challenges traditional IT organizations face when rolling out and managing Hadoop.

As for Meagan's BI tools, there's a difference between should and could with respect to allowing them to connect directly to Hadoop. Although most BI tools can connect to Hadoop, the environment is typically not optimized for analytics and queries.

As noted previously, this space is evolving and direct access to Hadoop-based analytics environments is becoming more prominent. We believe that the user experience and performance of traditional BI tools will be much better against an environment optimized for access versus low-cost storage.

Similarly, the same recommendation holds true for connecting mobile devices. Although being able to access information on the move is important to Southern Style, Hadoop is not currently optimized for data distribution to a mobile device, especially for the large amounts of data typically involved. It can be difficult to navigate on a mobile device and users could struggle to interpret the data.

On the topic of sandboxes, the answer is "it depends." It is important to acknowledge that Hadoop is one of the components of next-generation analytics environments (and there could be a business use case that dictates a Hadoop sandbox), but it is more likely a component of the discovery or exploratory environment and not the sandbox itself.

Furthermore, it is likely that the exploratory needs of the business will require more in terms of analytics abilities and access than native Hadoop can offer by itself.

Meagan and her team are on the right track recognizing that Hadoop is synonymous with big data, but there's much more to the story. They should think of it as the core to big insights. Hadoop enables companies to discover things that they didn't know and even take it a step further to "what they didn't know, they didn't know."

To launch Southern Style's foray into creating a next-generation analytics environment, Meagan should identify the top three most promising business cases for big data and initiate pilot projects that can be used to justify Southern Style's future investment in hardware and software resources.

JOEY D'ANTONI

Meagan is facing the challenge and opportunity many BI leaders face—an opportunity to reduce overall costs while being able to consume more data and provide more informed insights to her internal customers. Hadoop has presented IT organizations with a way to reduce storage and licensing costs while providing the scale needed to deal with Web and social media data.

In my opinion, the biggest challenge Meagan will face is that her team is has worked with "vanilla" business intelligence tools. Learning Hadoop and the other associated components could provide one hurdle to project success. In the Hadoop implementations I have worked with, getting staff comfortable with the toolset has been the biggest challenge. Many of these challenges are just getting IT infrastructure support given the differences in the platform. If your teams are not comfortable with Linux, you are likely going to have problems.

To reduce the risk of this effort, Meagan should launch an experiment to start working with Hadoop with a cloud provider such as Amazon's Elastic Map Reduce or Microsoft's HD Insight. The cloud overcomes many of the challenges many traditional IT organizations face when rolling out and managing Hadoop. This also lowers the cost of entry to the project and allows her staff to focus exclusively on the data.

A good entry point for this would be Web log and social media data, which already lives in the Internet, reducing challenges associated with migrating data into the cloud. At this point, Meagan should not try to migrate the entire data warehouse to Hadoop. Although Hadoop can augment a data warehouse, there are still many reasons (in Southern Style's case, stability) to maintain the traditional relational data warehouse.

> Hadoop allows teams to rapidly prototype and test new data sources in a fashion that's not practical in a traditional data warehouse.

As Southern Style's team gets more comfortable with Hadoop, it may make sense to start to transition some ETL processes to Hadoop. Many organizations have moved to using Hadoop as a data ingest engine (I'll refrain from the "data lake" buzzword). Classic ETL tools are very powerful, but they can also be quite expensive, and Hadoop offers the opportunity to scale out workloads at a lower cost and with a higher degree of flexibility than traditional tools.

Hadoop is a significantly more mature platform than it was just two or three years ago, and this also applies to vendor tools that work with it. Most major business intelligence tools that can connect to a data warehouse or OLAP cube can connect to modern platforms such as Hadoop and HBase. Southern Style needs to ensure they are running the latest versions of their BI software stack because these connectors are frequently in the latest tool releases.

A similar effect is seen with mobile solutions—Southern Style currently provides mobile BI to key executives. Although adding another data platform will make this challenging, most current mobile BI platforms provide support for big data sources.

One architectural consideration, particularly around mobile and with cloud-based data, is to perform as much initial aggregation as possible before pushing reports down to clients for two reasons: you want the users to have a good experience from a performance and response time perspective and downstream cloud traffic costs money (pre-aggregating minimizes these costs).

Perhaps the best use case of Hadoop in any data warehousing scenario is for prototyping and development. In the traditional data warehouse, 80 percent or more of the effort is associated with making sure the

ETL process is developed and is functioning correctly. Because Hadoop is schemaless and can manage data that's not as "clean" as a traditional relational database might require, Hadoop allows Meagan's team to rapidly prototype and test new data sources in a fashion that's not practical in a traditional data warehouse.

A

SRINIVAS VARANASI

The excitement of Meagan and her senior managers is justified in their move to store and analyze new big data sources to expand and sustain their business operations through an international network of stores. Though big data was no longer on Gartner's hype curve in July 2015, Meagan has to exercise a pragmatic approach in the fast-changing environment of technology choices, budget constraints, and limited supply of highly skilled resources.

A data lake based on Hadoop can only add to complexity unless an appropriate strategy is crafted linking the technology to business objectives, such as expansion of markets, channels, cross-sell/up-sell of new designs, monitoring customer behavior through social media (to enhance customer retention), and faster designs to market to satisfy customer tastes.

Hadoop-based Data Lake as Data Storage

A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, unstructured, and semi-structured data. Factors to consider include:

> Data lakes are a good fit for migration of ETL processes that consume processing cycles of enterprise data warehouses.

- A data lake is flexible; it can store both structured (traditional data sources) and big data (from such sources as social media, machine data, and web log data). A data lake can store both repetitive (traditional transactional, customer-service interactions data) and nonrepetitive data (Twitter and social media data).
- A data lake enables agility; their "schema-on-read" feature means nontraditional data can be merged and enriched with traditional data for monetization. Data exploration is more powerful when data models can be built on the fly.
- The evolution of SMAC technologies has led traditional and emerging vendors to offer Hadoop and traditional data solutions in the cloud.
- Data security is still evolving and appropriate security policies and infrastructure measures must be taken to protect customer privacy, corporate confidentiality, and financial data as well as support regulatory compliance.
- Data integrity and maintaining a single version of the truth is not possible in a Hadoop-based data lake.
- Hadoop offers a lower total cost of ownership and reduces user community training.

Hadoop for All ETL Processes

There is a view that a data lake can be a staging area for all ETL processes. This sort of scale-out ETL allows data to be distilled into a form that is loaded into a data warehouse for wider use. Data lakes are also a good fit for migration of ETL processes that consume processing cycles of enterprise data warehouses.

Scaling data ingestion into the data lake to match digital and emerging data sources is beneficial because ETL can take place within a data lake. A significant advantage is that the ETL process can run against data from enterprise applications and big data sources at the same time.

A large European bank attempted to transform their information architecture, including ETL, to a

Hadoop-based data lake. The bank discovered using HDFS (Hadoop Distributed File Systems) and MapReduce later is not sustainable and started considering use of Spark. As technologies evolve, it is imperative for Meagan to consider incremental changes instead of a big bang approach in migrating ETL processes to data lakes.

BI Tools' Connectivity to Hadoop

BI tools traditionally were designed for small volumes of structured data. Hadoop was designed for batch processing of complex data types at scale. In the absence of defined dimensions and clear facts, data analysts can identify trends to adapt to a new interface in a Hadoop cluster. BI-Hadoop integration is feasible with SQL/JDBC/ODBC. Hive's JDBC interface does not provide enough metadata information. Both MicroStrategy and Tableau are vendors certified for their latest versions on Hadoop. Use cases include:

- Visually explore subject-matter extract in-memory through a one-time query to Hadoop
- Any self-service parameterized queries directly to Hadoop
- Users can connect to Hadoop cluster and then extract data into Tableau's fast in-memory data engine without waiting for MapReduce to complete processing

I would advise Meagan to explore options based on business requirements to leverage the existing BI tools' connectivity to Hadoop.

Does Hadoop Fit In with Mobility?

Mobile applications generally are not designed as new Hadoop applications because Hadoop's value is greater for databases that are petabytes (or larger) in size than for typical databases used by mobile applications. Hadoop has significant setup and processing overhead and is less helpful in analytics processing performed iteratively, especially as several sequential dependent steps.

To make Hadoop useful for mobile applications, reverse engineer the applications and take advantage of Hadoop's benefits; pay particular attention to tasks that run no more than daily and handle data (unstructured or semistructured). The information can be represented as a series of traditional databases that are, effectively, abstractions of raw data.

Hadoop Sandbox

Sandboxes on Hadoop are 100 percent open source; the features and frameworks are free of cost. Meagan should encourage innovation labs to promote innovation within Southern Style. The new functionalities can be tested in sandboxes before migrating them to a production environment.

These innovations can be used for training the existing staff, especially introducing staff to languages such as Hive, Pig, and Python for exploratory analytics and to solve business problems. Sandboxes can host tutorials on Hadoop, Spark, Storm, HBase, Kafka, Hive, Ambari, and YARN built on experience gained from learners.

Meagan might consider sandboxes as a proof-of-concept environment to solve business issues. ■