pentaho
A Hitachi Group Company

Hadoop
and the
Analytic Data Pipeline

Hadoop and the Analytic Data Pipeline

TABLE OF CONTENTS

INTRODUCTION

Hadoop is disruptive.

Over the last five years, there have been few more disruptive forces in information technology than big data – and at the center of this trend is the Hadoop ecosystem. While everyone has a slightly different definition of big data, Hadoop is usually the first technology that comes to mind in big data discussions.

When organizations can effectively leverage Hadoop, putting to work frameworks like MapReduce, YARN, and Spark, the potential IT and business benefits can be particularly large. Over time, we've seen pioneering organizations achieve this type of success – and they've established some repeatable value-added use case patterns along the way.

Examples include optimizing data warehouses by offloading less frequently used data and heavy transformation workloads to Hadoop, as well as customer 360-degree view projects that blend operational data sources together with big data to create on-demand intelligence across key customer touch points. Organizations have achieved what can be best described as "order of magnitude" benefits in some of these scenarios, for instance:

- Reducing ETL and data onboarding process times from many hours to less than an hour
- Cutting millions of dollars in spending with traditional data warehouse vendors
- Accelerating time to identify fraudulent transactions or other customer behavior indicators by 10 times or more

> Hadoop is hard – but the right tools make it easier.

Given these potentially transformational results, you might ask – "Why isn't every organization doing this today?" One major reason is simply that Hadoop is hard. As with any technology that is just beginning to mature, barriers to entry are high. Specifically, some of the most common challenges to successfully implementing Hadoop for value-added analytics are:

- A mismatch between the complex coding and scripting skillsets required to work with Hadoop and the SQL-centric skillsets most organizations possess
- High cost of acquiring developers to work with Hadoop, coupled with the risk of having to interpret and manage their code if they leave
- Sheer amount of time and effort it takes to manually code, tune, and debug routines for Hadoop
- Challenges integrating Hadoop into enterprise data architectures and making it "play nice" with existing databases, applications, and other systems

These are some of the most readily apparent reasons why Hadoop projects may fail, leaving IT organizations disillusioned that the expected massive ROI (return on investment) has not been delivered. In fact, some experts are expecting the large majority of Hadoop projects to fall short of their business goals for these very reasons.[1]

1 "Through 2018, 70 percent of Hadoop deployments will not meet cost savings and revenue generation objectives due to skills and integration challenges," Gartner Analyst, Nick Heudecker; infoworld.com, Sept 2015.

The good news is that traditional data integration software providers have begun to update their tools to help ease the pain of Hadoop, letting ETL developers and data analysts integrate and process data in a Hadoop environment with their existing skills. However, leveraging existing ETL skill sets alleviates just one part of a much larger set of big data challenges.

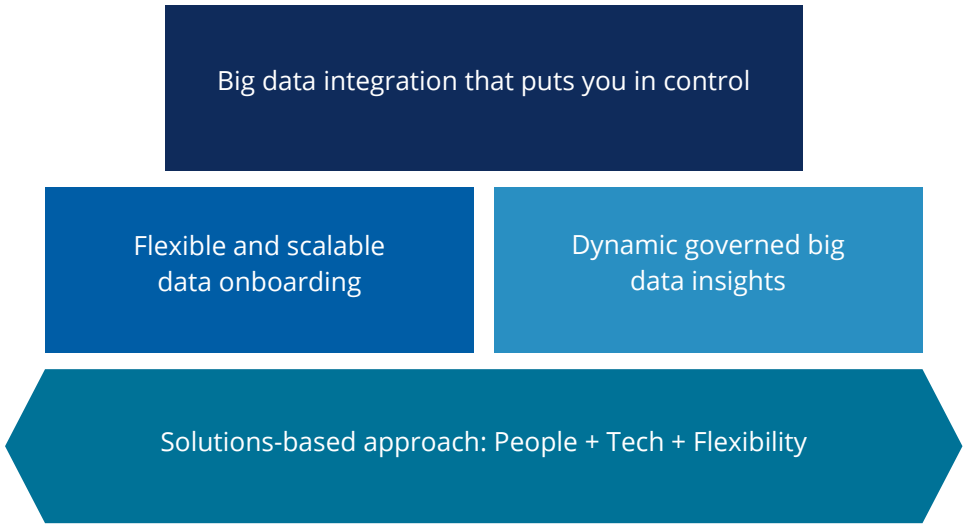
Taking a Holistic View of Big Data Projects

Stepping back, it is important to recognize that determining how to deliver business value from Hadoop still represents the second most frequent challenge for those considering the technology (after skills gaps).[2] Organizations need to keep this front of mind and avoid landing their Hadoop initiative in the "science project" or "experimental" category.

To be clear, user-friendly ETL tools for big data can certainly accelerate developer productivity in Hadoop use cases. However, if there isn't a clear plan that addresses the end-to-end delivery of integrated, governed data, as well as analytics to address business goals, a lot of the potential benefits of Hadoop will be left on the table. Organizations may achieve some moderate cost take-out benefits, but transformative business results will be much harder to achieve.

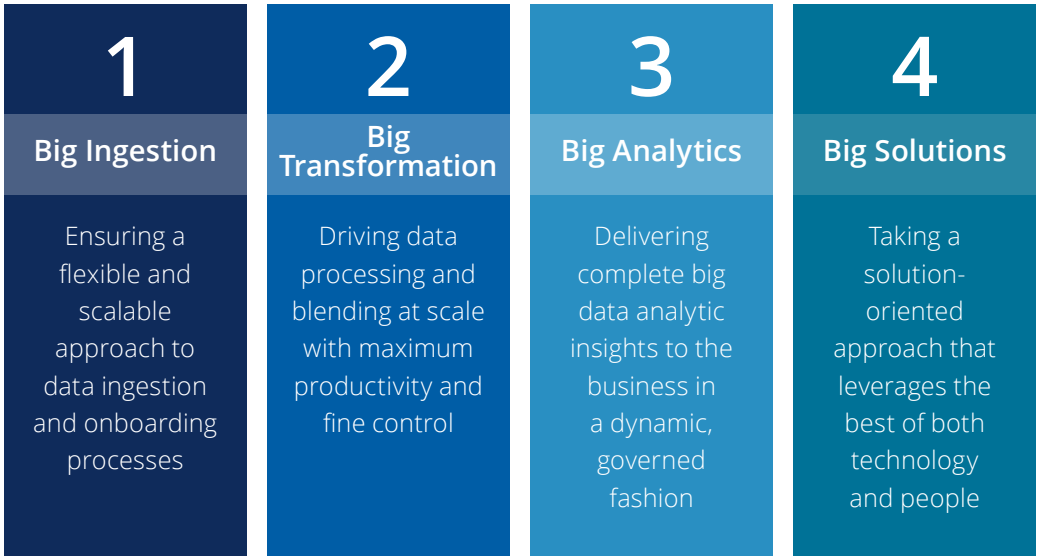
In order to maximize the ROI on their Hadoop investment, enterprises need to take a "full pipeline" view of their big data projects. This means approaching Hadoop in the context of end-to-end processes: starting at raw data sources, moving through data engineering and preparation inside and outside Hadoop, and finally leading to the delivery of analytic insights to various user roles, often as a part of existing business processes and applications.

A HOLISTIC APPROACH TO THE BIG DATA PIPELINE AND PROJECT LIFECYCLE



2 "Gartner Survey Highlights Challenges to Hadoop Adoption", gartner.com, May 2015.

Paying attention to the whole data analytic pipeline, including administration and orchestration of the overall process, keeps the IT group focused on the ultimate value it is delivering to the business as well as the broader impact on people, processes, and technology throughout the organization. In particular, there are four categories teams need to focus on in this comprehensive approach to Hadoop projects:



The remainder of this paper will dive into each of these categories in more detail.

Big Ingestion

ENSURING A FLEXIBLE AND SCALABLE APPROACH TO DATA INGESTION AND ONBOARDING PROCESSES

Naturally, the first step in an enterprise data pipeline involves the source systems and raw data that will ultimately be ingested, blended, and analyzed. Market experience dictates that the most important big data insights tend to come from combinations of diverse data that may initially be isolated in silos across the organization.

As such, a key need in Hadoop data and analytics projects is the ability to tap into a variety of different data sources, types, and formats. Further, organizations need to prepare not only for the data they want to integrate with Hadoop today, but also data that will need to be handled for potential additional use cases in the future.

The following types of data sources and formats are often part of Hadoop analytics projects:

- Data warehouses and RDBMs containing transactional customer profile data
- Log file and event data including web logs, application logs, and more
- Data in semi-structured formats including XML, JSON, and Avro
- Flat files, such as those in CSV format
- Data housed in NoSQL data stores, such as HBase, MongoDB, and Cassandra
- Data pulled from web-based APIs as well as FTP servers
- Cloud and on-premise application data, such as CRM and ERP data
- Analytic databases such as HPE Vertica, Amazon Redshift, and SAP HANA

> Scalability of data ingestion and onboarding processes is mission-critical

Organizations are also finding that cost and efficiency pressures as well as other factors are leading them to use cloud-computing environments more heavily. They may run Hadoop distributions and other data stores on cloud infrastructure, and as a result, may need data integration solutions to be cloud-friendly.

This can include running on the public cloud to take advantage of scalability and elasticity, private clouds with connectivity to on-premises data sources, as well as hybrid cloud environments. In a public cloud scenario, organizations may look to leverage storage, databases, and Hadoop distributions from an overall infrastructure provider (In the case of Amazon web services, this would mean S3 for storage, Amazon Redshift for analytic data warehousing, and Amazon Elastic MapReduce for Hadoop).

As Hadoop projects evolve from small pilots to departmental use cases and, eventually, enterprise shared service environments, scalability of the data ingestion and onboarding processes becomes mission-critical. More data sources are introduced over time, individual data sources change, and frequency of ingestion can vacillate. As this process extends out to a hundred data sources or more, which could even be a range of similar files in varying formats, maintaining the Hadoop data ingestion process can become especially painful.

At this point, organizations desperately need to reduce manual effort, potential for error, and amount of time spent on the care and feeding of Hadoop. They need to go beyond manually designing data ingestion workflows to establish a dynamic and reusable approach while also maintaining traceability and governance.

Being able to create dynamic data ingestion templates that apply metadata on-the-fly for each new or changed source is one solution to this problem. According to a recent best practices guide by Ralph Kimball, "consider using a metadata-driven codeless development environment to increase productivity and help insulate you from underlying technology changes."[3] Not surprisingly, the earlier organizations can anticipate these needs, the better.

What to look for from vendors:

DATA INGESTION AND ONBOARDING

- ☐ Easy connectivity to traditional data sources including data warehouses, flat files, and enterprise applications
- ☐ Straightforward connectivity to Hadoop, NoSQL stores, and support for a variety of semi-structured and non-relational data formats
- ☐ Ability to deploy platform in public, private, and hybrid cloud environments and take advantage of cloud-based big data
- ☐ Transformation templates that make it possible to generate jobs on the fly and scale data onboarding out to many more data sources with minimal manual effort

3 Ralph Kimball, Kimball Group, "Newly Emerging Best Practices for Big Data".

Big Transformation

DRIVING SCALABLE DATA PROCESSING AND BLENDING WITH MAXIMUM PRODUCTIVITY AND FINE CONTROL FOR DEVELOPERS AND DATA ANALYSTS

Once enterprises are able to successfully pull a variety of data into Hadoop in a flexible and scalable fashion, the next step involves processing, transforming and blending that data at scale on the Hadoop cluster. This enables complete analytics, taking all relevant data into account, whether structured, semi-structured, or unstructured.

As touched on earlier, it is essentially a "table stakes" requirement to leverage an intuitive and easy-to-use data integration product to design and execute these types of data integration workflows on the Hadoop cluster. Providing drag and drop Hadoop data integration tools to ETL developers and data analysts allows enterprises to avoid hiring expensive developers with Hadoop experience.

In a rapidly evolving big data world, IT departments also need to design and maintain data transformations without having to worry about changes to the underlying technology infrastructure. There needs to be a level of abstraction away from the underlying framework (whether Hadoop or something else), such that the development and maintenance of data-intensive applications can be democratized beyond a small group of expert coders.

> Provide drag and drop Hadoop data integration tools to ETL developers and data analysts

This is possible with the combination of a highly portable data transformation engine ("write once, run anywhere") and an intuitive graphical development environment for data integration and orchestration workflow. Ideally, this joint set of capabilities is encapsulated entirely within one software platform. Overall, this approach not only boosts IT productivity dramatically, but it also accelerates the delivery of actionable analytics to business decision makers.

Ease of installation and configuration is a related element that enterprises can look to in order to drive superior time to value in Hadoop data integration and analytics projects. This is fairly intuitive – the more adapters, node-by-node installations, and separate Hadoop component configurations required, the longer it will take to get up and running. However, underlying solution architecture, and by extension configuration processes, can have important additional operational implications.

For instance, as more node-by-node software is installed and more cluster variables are tuned, it is more likely that an approach will risk interfering with policies and rules set by Hadoop administrators. Also, more onerous and cluster-invasive platform installation requirements can create problems including:

- Repetitive manual installation interventions
- Increased risk to change and reduced solution agility
- Inability to work in a dynamic provisioning model
- Reduced architectural flexibility
- Lower cost effectiveness

> Be wary of "black box" approaches to data transformation on Hadoop

Organizations taking a holistic approach to Hadoop data analytics will look beyond simply insulating traditional ETL developers from the complexity of Hadoop to providing different roles with the additional control and performance they need. If a broader base of Hadoop developers, admins, and data scientists should be involved in the overall data pipeline, those roles need to be empowered to work productively with Hadoop as well.

Enterprises should be wary of "black box" approaches to data transformation on Hadoop, and instead, opt for an approach that combines ease of use and deeper control and visibility. This includes native, transparent transformation execution via MapReduce, direct control over spinning up or down cluster resources via YARN, ability to work with data in HBase, and integration with tools like Sqoop for bulk loads and Oozie for workflow management. It can also extend out to providing the ability to orchestrate and leverage pre-existing scripts (Java, Pig, Hive, etc.) that organizations may still want to use in conjunction with other visually designed jobs and transformations.

An alternative approach to big data integration involves the use of code generation tools, which output code that must then be separately run. In addition, because these tools generate code, that code is often maintained, tuned, and debugged directly – which can create additional overhead for Hadoop projects. Code generators may provide fine-grained control, but they normally have a much steeper learning curve. Use of such code-generators mandates iterative and repetitive access to highly skilled technical resources familiar with coding and programming. As such, total cost of ownership (TCO) should be carefully evaluated.

What to look for from vendors:

DATA TRANSFORMATION AND BLENDING AT SCALE

- ☐ Intuitive drag and drop design paradigm for big data jobs and transformations, with ability to configure as needed
- ☐ Data integration run-time engine that is highly portable across different data storage and processing frameworks, drastically reducing need to re-factor data workflows
- ☐ Fast, repeatable configuration to run data transformations on Hadoop that minimizes node-by-node installation and cluster invasiveness
- ☐ Native and scalable ability to execute data transformations as Hadoop MapReduce jobs in-cluster
- ☐ Broad, transparent Hadoop ecosystem integration including YARN (job resource management), HBase (NoSQL store), Sqoop (bulk load), Oozie (workflow management), existing Pig scripts and more
- ☐ Encapsulation of all functionality within the data integration and analytics software, with no need to generate and manage separate code

Big Analytics

DELIVERING COMPLETE ANALYTIC INSIGHTS TO THE BUSINESS IN A DYNAMIC, GOVERNED FASHION

A prerequisite to unlocking maximum analytic value from Hadoop is carefully considering all relevant business end users, as well as business processes and applications (internal and external) that the project should touch. Different data consumers may need different tooling and approaches, depending on their needs and levels of sophistication.

As data scientists and advanced analysts begin to query and explore blended data sets in Hadoop, they will often make use of data warehouse and SQL-like layers on Hadoop, such as Hive and Impala. Thanks to a familiar type of query language, these tools do not take long to learn. As such, skilled data analysts should seek out data integration and analytics platforms that provide operational reporting and visual analytics directly on Hive and Impala.

At the same time, it's important to note that SQL layers on Hadoop do present limitations in several ways. First, they may not provide the degree of interactivity expected in today's reporting and analytics tools (when used on relational data sources). In particular, there may be latency limitations related to the complexity of queries and amounts of data involved.

> Hadoop as part of the broader analytic pipeline is crucial.

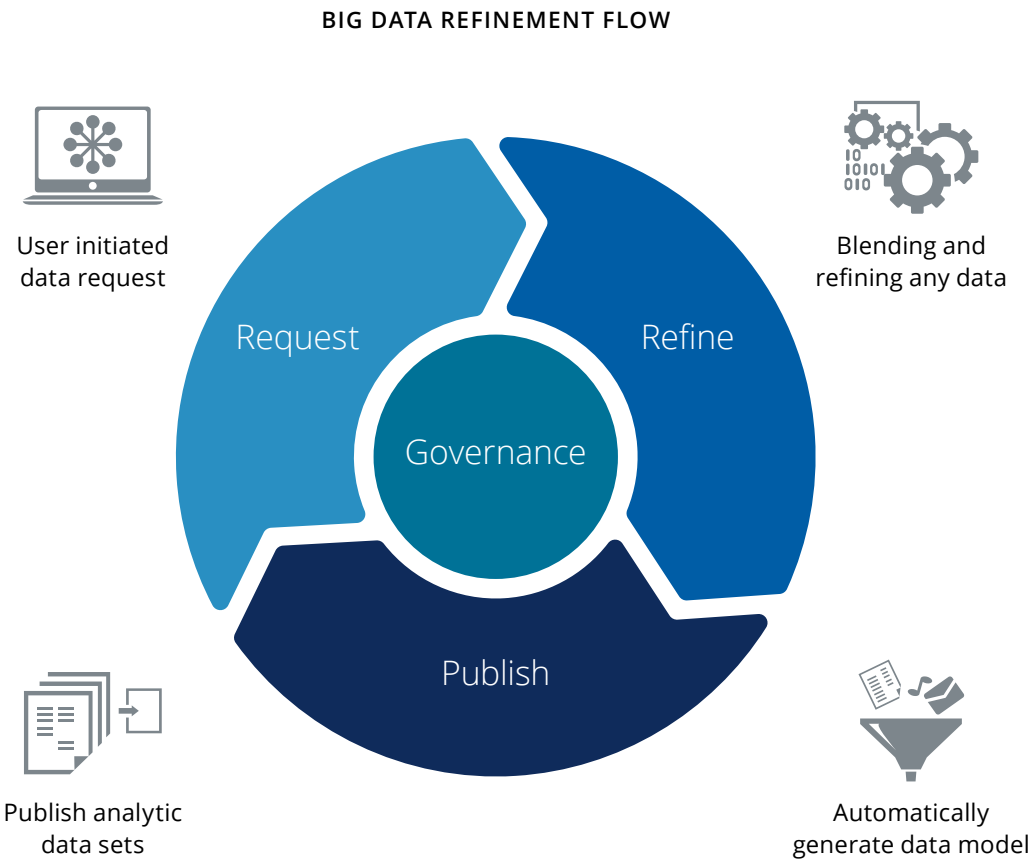
While this is acceptable in the analytics prototyping phase, the performance and usability are unlikely to satisfy the requirements of larger groups of analysts and business users in production environments. The wrong query at the wrong time can potentially strain Hadoop cluster resources, interfering with the completion of other integration processes.

Enterprises are used to providing business users with analytics tools that sit on top of highly governed, pre-processed data warehouses. The data is, for the most part, trusted and accurate, while pre-built analytic cubes offer fast answers to the business questions that business users may want to ask of the data. Conversely, in the world of Hadoop, it is a much greater challenge to provide direct analytics access at scale that is both highly governed as well as easily and interactively consumed by the analytics end user. In many cases, there may be so much data in the Hadoop cluster that it may not even make sense to pre-process it as in a data warehouse scenario.

This is another circumstance where considering Hadoop as part of the broader analytic pipeline is crucial. Specifically, many organizations are already familiar with high performance relational databases that are optimized for interactive end-user analytics – or "analytic databases." Enterprises are finding that a highly effective way to unleash the analytic power of Hadoop is to deliver refined data sets from Hadoop to these databases.

The most effective approach enables the business user or analyst to intuitively request the subset of Hadoop data he or she would like to analyze. A user selection can trigger on-demand data processing and blending in the Hadoop environment, followed by delivery of an analytics-ready data set to the end user for ad hoc analysis and visualization.

An illustration of this process flow is depicted below, starting at the upper left:



Given the complexities of working with Hadoop, it is especially important for the IT organization to be able to architect this process to establish the same level of trust it already has with respect to the group's enterprise data warehouse. However, once the process is established, business requests to IT for Hadoop data sets is drastically reduced. Of course, it is also important to provide an intuitive end user interface for requesting and exploring these on-demand data marts.

Finally, it is key to integrate and operationalize advanced predictive and statistical modeling into the broader big data pipeline. Despite their potential to create path-breaking insights, data scientists often find themselves outside the broader enterprise data integration and analytics production process. Further, the majority of time and effort in a predictive analytics task is often spent preparing the data, rather than actually analyzing and modeling it.

The more the data integration and analytics approach enables collaboration between data scientists and the broader IT team, the quicker it will be to develop and implement new models for forecasting, scoring, and more – leading to faster business benefits. In particular, time to insight can be accelerated by allowing data scientists to develop models in their framework of choice (R, Python, etc.) and apply those models directly within the data transformation and preparation workflow. These models can then be more easily embedded in regularly occurring business processes.

What to look for from vendors:

ANALYTICS FOR GOVERNED, DYNAMIC INSIGHTS

- ☐ Reporting and visual analysis tools that work on data in Hive and Impala
- ☐ Ability to orchestrate end user driven data refinement, blending, modeling, and delivery of data sets in a big data environment, including Hadoop and analytic databases
- ☐ Provisioning of intuitive business user interfaces to select and analyze data in a big data environment
- ☐ End to end visibility into and trust in the data integration and analytics process, from raw data to visualizations on the glass
- ☐ Ability to integrate existing predictive models from R, Python, and other advanced analytics frameworks into the data preparation workflow

Big Solutions

TAKING A SOLUTION-ORIENTED APPROACH THAT LEVERAGES THE BEST OF BOTH TECHNOLOGY AND PEOPLE

While many advancements have been made in the Hadoop ecosystem over the last several years, Hadoop is still maturing as a platform for use in production enterprise deployments. Moreover, anyone who's been involved with enterprise technology initiatives knows that requirements evolve and tend to be "works in process" more than set in stone. Hadoop represents a major new element in the broader data pipeline, and related initiatives usually require a phased approach.

As a result, software evaluators will not find one off-the-shelf tool that satisfies all current and forward-looking Hadoop data and analytics requirements as is. "Future proofing" is in danger of becoming an overused word in conversations around big data today, but flexibility and extensibility should be part of all project checklists. Ralph Kimball elaborates on this set of needs in more detail:

> "...plan for disruptive changes coming from every direction: new data types, competitive challenges, programming approaches, hardware, networking technology, and services offered by literally hundreds of new big data providers...
>
> ...maintain a balance among several implementation approaches including Hadoop, traditional grid computing, pushdown optimization in an RDBMS, on-premise computing, cloud computing, and even your mainframe." [4]

The ability to port transformations to run seamlessly across different Hadoop distributions is a starting point, as many organizations are not sure what their enterprise standard distribution will be down the road. However, true durability requires an overall platform approach to flexibility that aligns with the open innovation that has driven the Hadoop ecosystem, including:

"Future proofing" is in danger of becoming an overused word.

- Open architectures based on open standards that are easy for IT teams to understand
- Ability to easily leverage existing scripts and code across a variety of frameworks, whether that means Java, Pig scripts, Python, or others
- Open APIs and well-defined SDKs that facilitate solution extensions to introduce add-on data and analytics capabilities for specific use cases
- Seamless ability to embed reports, visualizations, and other analytic content into existing business applications and processes

In addition, it takes more than just the right technology platform: the ability to leverage the right people is arguably more important to project success. Too often, organizations experience delays and underwhelming results when it comes to Hadoop data integration and analytics. The problem isn't always with the underlying technology – rather it is very common that best practices solution architectures and implementation approaches are not being followed. Working with a seasoned partner with deep expertise in Hadoop data and analytics projects can help set teams on the right path from the start and avoid costly course corrections (or worse) later on.

4 Ralph Kimball, Kimball Group, "Newly Emerging Best Practices for Big Data".

Since Hadoop itself is so new, IT teams should place a premium on a data integration and analytics provider's track record of customer success with Hadoop-specific projects, not just generic data integration and analytics projects. In addition to vendor service offerings, organizations should consider the experience and expertise of the big data services team members. These should span the entire project lifecycle, from solution visioning and implementation workshops to in-depth training programs, architect-level support, and technical account management.

What to look for from vendors:

SOLUTIONS THAT LEVERAGE THE BEST OF PEOPLE AND TECHNOLOGY

- ☐ Portability of Hadoop data transformations to run across different commercial distributions with minimal overhead
- ☐ Open platform architecture based on open standards, as well as the ability to leverage existing scripts in languages like Java, Python, Pig, R and others
- ☐ Open APIs and well-defined SDKs that let users easily create platform extensions for new use cases
- ☐ Seamless ability to embed reports, visualizations, and other analytics into existing business applications and processes
- ☐ A well-established track record of customer success with big data and Hadoop projects, including multiple specific reference customers
- ☐ An experienced big data services organization, with offerings covering the full project lifecycle, from workshops and training to architect-level support and ongoing consulting services

Conclusion

Big data has the potential to solve big problems and create transformational business benefits. While a whole ecosystem of tools have sprung up around Hadoop to handle and analyze data, many of them are specialized to just one part of a larger process. In order to fulfill the promise of Hadoop, organizations need to step back and take an end-to-end view of their analytic data pipelines.

This means considering every phase of the process – from data ingestion to data transformation to end analytic consumption, and even beyond to other applications and systems where analytics must be embedded. It means not only tackling the tactical challenges like closing the big data development skills gap but also clearly determining how Hadoop and big data will create value for the business. Whether this happens through cost savings, revenue generation, better customer experiences or other objectives, taking an end-to-end view of the data pipeline will help promote project success and enhanced IT collaboration with the business.

In summary, organizations should keep the follow tenets of successful big data projects top of mind:

1. Ensuring a flexible and scalable approach to data ingestion and onboarding processes
2. Driving data processing and blending at scale with maximum productivity and fine control
3. Delivering complete big data analytic insights to the business in a dynamic, governed fashion
4. Taking a solution-oriented approach that leverages the best in both technology and people

What's next

Ready to get serious about boosting the analytics experience for your users? Check out these helpful resources.

- Learn how to architect end-to-end data management solutions with Apache Hadoop.
- Read a free excerpt from the Field Guide to Hadoop to learn about Hadoop's core technologies.
- See how Pentaho tackles Hadoop challenges head-on, from data integration to proven big data implementation patterns and expertise.
- See a quick demo of how Pentaho makes it easy to transform and blend data at scale on Hadoop without coding.

About Pentaho

A Hitachi Group Company

Pentaho, a Hitachi Group company, is a leading data integration and business analytics company with an enterprise-class, open source-based platform for diverse big data deployments. Pentaho's unified data integration and analytics platform is comprehensive, completely embeddable and delivers governed data to power any analytics in any environment. Pentaho's mission is to help organizations across multiple industries harness the value from all their data, including big data and IoT, enabling them to find new revenue streams, operate more efficiently, deliver outstanding service and minimize risk. Pentaho has over 15,000 product deployments and 1,500 commercial customers today including ABN-AMRO Clearing, BT, Caterpillar Marine Asset Intelligence, EMC, Landmark Halliburton, Moody's, NASDAQ and Sears Holding Corporation. For more information visit www.pentaho.com.

pentaho®
A Hitachi Group Company

Be social with Pentaho: