Eight Essential Checklists

for Managing the Analytic Data Pipeline

By Hitachi Vantara

March 2019

Contents

Introduction

Organizations of every type are racing to use their data better. We all know the importance of analytics and business intelligence (BI) when it comes to successfully operating a company. But in an age when the number of data sources and data volumes is exploding, it's essential to make certain that your data is analytics-ready at the beginning of the data pipeline, as opposed to random points where business users may need to use it.

Failing to take a holistic approach to your data pipeline can yield dark, unused data, or worse: It may compel organizations to make critical business decisions based on inaccurate data. Within any analytics pipeline, the right data management processes are paramount to providing accurate and reliable information. These processes help to future-proof your pipeline against changing analytic needs and emerging technologies.

451 Research, an IT research company, validated the importance of data management in their recent report titled "Data Platforms and Analytics Market Map 2018." The authors note that "Data management is an essential part of the analytics process, and is defined as the management of data using a number of specific tools – or a broader platform combining multiple tools – with the endgame of enabling analytics."

In other words, it's important to consider whether the right approach for your company is to try to assemble a patchwork of data management tools, one for each part of the data management process, or if time and cost savings can be obtained through a comprehensive platform.

This guide provides a checklist of eight essential categories to consider as you evaluate your vendor options, and flags potential pitfalls to guard against as you plan your pipeline. Each category is critical part of the analytic data pipeline, from data connectivity and preparation, to analytics (see Figure 1). You must also consider ease of use, flexibility and governance to ensure that the right people can access the right data at the right time.

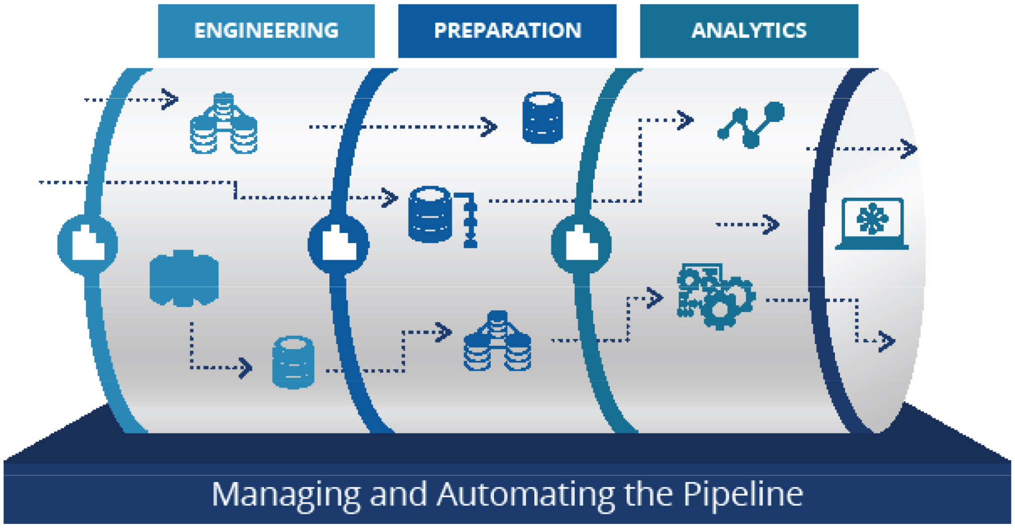

Figure 1. Essential Elements of Your Data Pipeline

Checklist 1
Data Connectivity

Data connectivity is essential to ensure that business-critical data is available for analysis and that your platform is ready to handle newer data sources and types – including big data sources – from which competitive advantage is so often gained.

To manage your data pipeline effectively, your tools must have the right connectivity to both traditional and emerging sources of structured, semi-structured and unstructured data. When evaluating potential vendors, it's important to ask questions about their connectivity capabilities when it comes to different types of data sources.

Data Source Types	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Flat Files	Can you access flat files saved via a variety of operating systems using a range of file storage options?	Vendors may support only Linux or Microsoft Windows operating systems. Also, vendors may not support solid-state drives (SSDs), flash drives, and other high-performance storage options.
Relational Databases	Do you connect to a wide variety of relational database management systems (DBMSs) via Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC)?	Vendors may support only popular databases such as Oracle DB, Microsoft SQL Server, IBM® DB2®, SAP Sybase or IBM Informix®. They may not support newer databases such as MySQL, PostgreSQL, Hypersonic SQL and H2.
Legacy or Nonrelational DBMSs	Do you read raw data from legacy or nonrelational data sources?	Vendors may connect to IBM AS/400 but not to mainframes, COBOL environments and other legacy systems.
Enterprise Resource Planning (ERP), Customer Relationship Management and Supply Chain Management (SCM) Sources	Can you connect to packaged applications?	Vendors may connect to popular systems such as SAP ERP and Salesforce CRM but not to newer systems such as OpenERP or Splunk.
Cloud Sources	Can you connect to cloud data sources?	Vendors may connect to only a subset of Amazon's data sources, which includes Amazon Redshift, Amazon S3, Amazon EMR, Amazon RDS, and Amazon DynamoDB.
Software-as-a-Service (SaaS) Applications	Can you connect to a variety of SaaS applications?	Vendors may connect to a few SaaS applications, such as Google Analytics or JIRA.
Application Programming Interface (API) Inputs	Can you use a variety of APIs to connect to different web services?	Vendors may support only web services such as REST API but not older file transfer protocols such as FTP or HTTP.
Industry Standards	Can you read a variety of industry-standard data streams and feeds?	Vendors may support standard formats for a specific vertical, such as SWIFT for financial services, but not older standards, such as EDI, or standards in a different vertical, such as HL7 in healthcare.

Continued on next page

Checklist 1

Data Connectivity (cont.)

Data Source Types	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Message Queues and Enterprise Application Integration (EAI) Products	Can you read messages from a variety of queues?	Vendors may support only Java Message Service (JMS) or IBM MQ server.
Semistructured Data (JSON, XML)	Do you read and process complex XML files?	Vendors may have limitations using XPath specifications, adding XML tags into a stream of data, or reading data from RSS and Atom feeds.
Log Files	Do you integrate and parse complex application logs?	Vendors may use manual coding for parsing infrastructure log files or mobile log files.
Social Data	Do you connect to prominent social media sites?	Vendors may have generic REST APIs and may not be optimized to read from Facebook or Twitter feeds.
Mobile Platforms	Do you collect mobile device information?	Vendors may be limited to Android devices from top manufacturers, such as Samsung or LG, and may not support closed systems, such as Apple iOS or Microsoft Windows Phone.
Spatial Data	Do you connect to Esri data and do you offer any geographic information system (GIS) features?	Vendors may not offer geolocation data or provide geocoding services.
Structured and Unstructured Data	Do you connect to a variety of structured and unstructured data sources?	Vendors may support a limited set of file types, such as PDF and RTF, and may not support audio image, and video formats. The vendor may access email through either POP or IMAP.
Web Clickstream Data	Do you connect to and parse clickstream data?	While vendors may be able to access server log files, they may not be able to access application logs that are often stored in RDBMSs, NoSQL, or other stores. Also, vendors may not provide connectivity to Google Analytics for clickstream data.
Call Detail Records (CDR)	Do you parse CDR data in different common message formats and protocols?	Vendors may not easily adapt to changes in CDR structure and may support only a basic structure.
Big Data Future-Proofing	Can you help provide future-proofing by connecting to different sources of big data?	While vendors may be able to connect to some NoSQL databases, such as Apache Cassandra or MongoDB, they might not be able to integrate data from different on-premises and cloud Hadoop distributions.

Checklist 2
Data Engineering

Data engineering requires more than just connecting to or loading data. Rather, it involves managing a changing array of data sources, establishing repeatable processes at scale, and maintaining control and governance. Whether an organization is implementing an ongoing process for ingesting hundreds of data sources into Hadoop or enabling business users to upload diverse data without IT assistance, onboarding projects tend to create major obstacles. Additionally, as technology researcher O'Reilly Media has noted, data engineering best practices mean that your data pipeline should be reproducible, consistent and productionizable. Consider vendors' components and processes that will enable you to go from information delivery to data integration and all the way through to analytics and reporting.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Drag-and-Drop User Interface (UI)	Do you offer a drag-and-drop capability?	Drag-and-drop steps may be too generic, requiring additional configuration and manual coding.
Repeatability	Do you source data in a repeatable fashion?	Vendors may need to tweak the ingestion process for even small changes in source structure.
Event Framework	Do you deliver data through automated processes initiated by a variety of events?	Vendors may not support event processing to automatically initiate the transformation process.
Operationalizing Self-Service Data Prep	Do you operationalize data preparation rules created by business users so they can be rerun on schedule or on demand?	Vendors may not understand the preparation rules since these may have been created by a different tool from another vendor.
Metadata Driven	Can you share and reuse data definitions and metadata?	Vendors may require that each step in the transformation process be hard-coded rather than metadata driven.
Integration	Do you permanently join data from different sources?	Vendors may support only a limited set of joins or joins of similarly structured data.
Prebuilt Data Integration (DI) Steps	Do you offer me a range of sort, lookup and join steps?	Vendors may offer only the most common variations of these steps.
Predictive Model Support	Do you operationalize advanced analytics models?	Vendors may require users to manually feed data to their models with a custom script at a later stage in the pipeline.
Clustering	Do you enable me to cluster servers to improve performance?	Vendors may limit clustering to a homogeneous set of servers having the same operating systems.
Load Balancing	Do you enable me to distribute the transformation process across different nodes?	Vendors' ability to distribute the transformation process may not consider the current load on the system outside the data engineering process.

Continued on next page

Checklist 2

Data Engineering (cont.)

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
High Availability With Automatic Restarts	Do you automatically recover processes from external failures?	Vendors may provide only partial support by restarting the engineering process from the beginning rather than the point of failure.
Data Federation	Do you extract data from multiple sources, integrate the data, and flow that data directly into reports?	Vendors may need to create a staging area for a BI tool to read from directly.
Cloud Processing	Can you process data using a cloud-processing capability?	Does the list of major cloud infrastructure providers include Amazon EC2 instances only?

Checklist 3

Data Delivery

Getting data where it needs to go is essential. Some solutions perform better with traditional data warehouses and some perform better with newer technologies. It's important to consider how to future-proof your solution so that you can avoid getting stuck with outdated technology when innovative companies and the open source community develop something new, requiring you to adapt to the changing data landscape. Here are some questions to consider to help you ensure your data pipeline strategy is flexible and nimble.

Locations To Deliver Data	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Data Marts	Do you connect to a variety of data marts?	Vendors may support only MySQLor PostgreSQL and may not support popular options such as Microsoft SQL Server.
Enterprise Data Warehouses (EDWs)	Do you connect to a variety of data warehouses and support bulk data movement?	Vendors may support only IBM Netezza, Teradata and possibly Oracle Exadata, and may not support bulk loading for newer data warehouses, such as SAP HANA and Greenplum Database.
Analytic Databases	Do you connect to a wide variety of analytic databases that use high-performance capabilities, such as columnar storage and MPP?	Vendors may support on-premises analytic databases, such as Infobright, Greenplum Database and HP Vertica, while skipping cloud databases, such as Amazon Redshift.
NoSQL Databases	Do you connect to schemaless NoSQL stores?	Vendors may support MongoDB and Apache Hbase while ignoring Apache Cassandra and Apache CouchDB.
Hadoop	Do you connect to and run transformations natively on a variety of Hadoop distributions and connect with the major components of the Hadoop ecosystem?	Vendors may connect to and run transformations natively as MapReduce or Spark processes on Cloudera and Hortonworks, but they may not support other popular distributions, such as MapR and Amazon EMR. Vendors may be unable to connect with Hive or Impala. They may fail to integrate with other Hadoop ecosystem components, such as Hbase, Oozie, Sqoop and YARN. They may not support Avro file formats.
In-Memory Data Sources	Do you connect to and take advantage of in-memory databases?	Vendors may support popular SAP HANA databases but not others, such as EXASOL, H2 and Infobright.

Checklist 4

Data Preparation

As Forbes noted in 2016, data scientists spend up to 80% of their time simply preparing data – time that could be better spent building analytical models[1]. Standalone tools that help with data preparation may lack the flexibility to blend both traditional and new unstructured data sources. The more standalone vendors you use, the more likely you are to run into problems in going from one stage of the data pipeline to another. Here are some considerations when choosing the right vendor to help you with an end-to-end data pipeline solution.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Data Discovery	Do you allow me to easily search, explore and discover data sources, tables, columns and files and to request permission if needed?	Vendors may allow you to discover data sources without considering your security needs.
Data Access	Is the connectivity for self-service users the same as it would be for IT or is it limited to a subset of data sources?	IT assistance may be needed to connect to many data sources, making it impossible to be a self-service tool.
Access Automation	Do you provide capabilities to save and automate access to data?	Access could be very narrowly defined, requiring the process to be repeated for each set of sources.
Visual Examination	Do you offer me tools to visually examine data from different sources to determine fitness for purpose? Is this view limited to the source or to all the steps along the transformation pipeline? Is visualization native to the tool or is it integrated with an analytics tool?	Vendors may provide visibility to just the source or target data, and may have no visibility to the intermediate transformation steps of the data analytics pipeline.
Data Profiling and Sampling	Do you offer a few quality metrics and statistical analysis?	Vendors may fail to highlight missing data. Statistics may be available for only a small subset of the data.
Data Relationships	Do you create relationships among multiple data sources?	Vendors may offer relationships only for structured data sources and not for other types of data.
Data Cleansing	Do you provide an easy way to resolve errors in data? Vendors may not provide smart text-processing algorithms that add users' actions to resolve data errors to their list of available actions?	Vendors may not provide smart text-processing algorithms that add users' actions to resolve data errors to their list of available actions."
Definitions	Do you offer a means to label data sets with additional context?	Vendors may limit definitions to technical definitions and may not include broader business definitions.

Continued on next page

[1] Gil Pres, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," March 23, 2016, Forbes.com

Checklist 4

Data Preparation (cont.)

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Business Glossary	Do you offer a glossary of predefined business terms that could be expanded?	Vendors may offer a limited business glossary that cannot be expanded with customized definitions.
Collaboration	Do you offer capabilities to capture tribal knowledge?	Vendors may not offer the ability to share content and commentary with other users.
Data Blending and Enrichment	Do you offer the ability to blend multiple data streams? Can you blend both traditional data and semistructured data? Do you blend dozens of data sources efficiently? Can data blending be accomplished with zero coding?	Vendors may be able to offer prototyping capabilities but may not be able to operationalize them. Vendors may not easily handle semistructured or unstructured data. Vendors may have a proprietary modeling language to blend data that takes time to learn.
Data Shaping	Do you create calculated fields, dimensions or aggregations?	Vendors may not offer templates for data shaping rules.
Preparation Automation	Do you templatize the entire preparation process?	Vendors may not offer design templates to automate data preparation.
Governance	Do you integrate with other parts of the analytics pipeline to provide a seamless experience?	Point vendors may not offer seamless integration, leading to different definitions, inaccurate data and loss of security and centralized control.
Advanced Analytics	Do you use prepared test data to improve the predictive component of your analytics models?	Most vendors are not able to incorporate any predictive analytics into their tools.

Checklist 5

Analytics

As your business needs evolve, it's important to have a platform that can evolve with you. Vendors that provide a fixed library of analytics options may not have the flexibility you need. Being able to leverage the best of predictive analytics and to embed analytics into your existing business processes or even in your software are critical for getting the most business value.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Reporting	Do you provide traditional linear-style reporting tools? Can business users create reports or is this function limited to administrators?	Many vendors don't offer traditional reporting tools, instead choosing more attractive visualization tools.
Dashboards	Do you offer numerous templates for end users to drop into a variety of visualizations and reports? Are changes made in one section of the dashboard reflected in other sections?	Many vendors limit the variety of interactive and analytic components that can be part of the dashboards. Different dashboard sections may not automatically synchronize after changes in one section.
Visualization	Do you offer a variety of bar or column charts, heat grids, geo maps and scatter plots? Do you offer the ability to create custom charts and templates?	Many vendors limit geo-mapping layers, with no extensibility that they support out of the box. Vendors may limit customization to their proprietary visualization libraries rather than provide access to open libraries.
Ad Hoc Analysis	Do you offer ad hoc analysis via a web interface in addition to a desktop interface?	Some vendors limit ad hoc analysis to desktop tools rather than web interfaces.
Embedding	Do you offer the ability to custom brand the analytics interface? Do you offer open APIs to easily embed analytics offering without extensive coding? Can the back-end data integration offering also be embedded along with the analytics?	Vendors may limit custom branding abilities or offer proprietary APIs. Vendors may need to partner with a data integration solution provider that may have a different set of APIs, leading to greater complexity.
Virtual Data Sets	Do you analyze virtual data sets?	Vendors may require data to be staged in a physical table.

Checklist 6

Pipeline Automation and Management

As the saying goes, “Excel ETL (extract, transform, load)” isn’t scalable, and the methodology one person follows may not be followed by colleagues in other business units, which leads to nonstandard reporting. It’s vital to be able to automate as much of your data pipeline as possible so you can make the most of your team’s resources. Consider the following questions to help you choose a vendor that can speed up the pipeline from raw data to analytics and business insights.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Monitoring	Do you show step-by-step performance metrics to identify bottlenecks?	Vendors may limit monitoring to just execution history.
Auditing	Do you offer auditing capabilities to analyze usage and trends and plan capacity?	Vendors may not offer an easy way to audit usage history. Therefore, users would have to have to rely on their homegrown metrics with questionable accuracy to plan capacity.
Automation	Can a user select a custom data set that is facilitated by automated data blending?	Vendors may not be able to initiate data integration jobs based on data set choices selected from the front end.
Error Handling	Are you capable of diverting rows that are in error rather than bringing the entire transformation process to a hard stop?	Vendors that do not separate the transformation from the orchestration typically must use a hard stop to handle any errors.
Stream Processing	Can you execute transformations via microbatches to process streaming data?	Vendors may perform only batch processing.
Native Integration	Do you offer end-to-end capabilities from data ingestion to transformation to analytics? If not, are you seamlessly integrated with your partners?	Lack of tight integration within a product portfolio or across partners’ products leads to enhanced complexity and longer debugging time.

Checklist 7

Governance and Security

Data governance and security are not optional, and it's best to have a security plan rather than handle damage control after a breach. If you're working in a regulated industry, it's especially important to use a data pipeline platform that captures the flow of who did what, with what data, and when. As you're evaluating vendors, review their capabilities when it comes to data governance and security.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Security	Do you integrate with security providers, such as LDAP, single sign-on (SSO) or Microsoft Active Directory? Do you then further customize their security settings?	Vendors may have their own security framework, with no ability to integrate with enterprise security frameworks.
Data Curation	Do you offer the ability to manage the data throughout its life cycle?	Vendors may not offer the ability to mark data as old or to be archived.
Data Lineage	Do you offer the capability to track where the data came from and how it was prepared? Does the lineage transition across different stages, from ingestion to transformation and later to the visualizations and analysis?	Vendors may offer lineage to only their portion of the analytics pipeline, resulting in a loss of history.
Data Protection	Do you have the capability to apply regulatory policies and rules to protect sensitive data? Do you promote data sanctioned by governance?	Vendors who cater to individuals or departments typically lack the ability to discern between sensitive data and broadly shared data.
Multitenancy	Do you deliver the correct data, reporting content and UI to the appropriate group in a shared, cloud-based infrastructure?	Vendors may only offer one customization when it comes to data, content or UI, resulting in an incomplete solution.

Checklist 8

Extensibility and Scalability

The big data ecosystem surrounding Apache Hadoop includes dozens of tools, which are each constantly evolving. Much of the innovation in the last few years around data management, especially with big data, has taken place in the open source community. Accordingly, if you're considering a vendor based on proprietary rather than open source code, you might get left behind when tools evolve. To stay flexible, consider a vendor's extensibility and scalability features.

Capability	Questions To Ask Your Vendor	Potential Vendor Pitfalls
Marketplace	Do you offer the ability to extend the platform with a marketplace of transformation plugins or custom visualizations?	Proprietary "black box" vendors mean you are dependent on their technical expertise rather than open to new developments as technology evolves.
Analytics Scalability	Can you accommodate more than a handful of users? Can there be a combination of internal and external users?	Vendors may not scale beyond a department level, resulting in a sharp drop-off in system responsiveness if many users are accessing it simultaneously.
Data Scalability	Do you transform millions of rows of data by leveraging the existing cluster of servers used for storing the data?	Vendors using enterprise service bus architecture are only good for running small pipes of small packets of data. Vendors may not natively run their processes in storage server clusters and may require additional hardware.
In-Memory Processing	Do you have the flexibility to process data in-memory for speed or push it down to disk when data size increases?	Vendors that are capable only of in-memory data processing hit obstacles to scalability because data must be loaded before processing, which becomes difficult as data sizes increase.
Clustering	Do you take advantage of a dynamic provisioning of nodes to a cluster?	Vendors deployed in the cloud may require node-by-node installs, and need more maintenance by IT.
Load Balancing	Do you offer load balancing across several servers forwarding traffic in a round-robin fashion, worker server quotas or some other method?	Vendors without a built-in load balancer need to rely on external load balancers that don't understand the transformation process.
High Availability	Do you provide uninterrupted access to critical data and reduce errors in data delivery?	Vendors without these capabilities require all transformation processes to be restarted when an error occurs.

Hitachi Vantara's Pentaho suite of software provides various analytics solutions that can help you manage your analytic data pipeline. Learn more about Pentaho analytics.

About Hitachi Vantara

Data is your greatest asset, if you know how to use it. It reveals your path to innovation and outcomes that matter for business and society. Hitachi Vantara combines 100 years of OT and 60 years of IT experience to help data-driven leaders unlock the value in their data. Our unique Stairway to Value model uses machine learning and artificial intelligence to deliver tangible benefits driven by your data. We help you store, enrich, activate and monetize your data to improve customer experiences, create new revenue streams and lower costs. We listen. We understand. We work with you.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
HitachiVantara.com | community.HitachiVantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
HitachiVantara.com/contact