Datameer
stay curious

Big Data Analytics Buyer's Guide

Why Do I Need Big Data Analytics?

Big data analytics is the process of running sophisticated analytic techniques against very large and diverse data sets. This resulting value of big data analytics comes from new insights that make a company data-driven, and power critical business processes to build closer customer relationships, increase revenue, operate more efficiently and reduce risk.

Why Do I Need Big Data Analytics?

Big data analytics is a combination of complex data — large volumes of diverse data — and sophisticated analysis on this data. A big data analytic problem is defined by having both of these characteristics present.

Analytic Complexity

The analytic techniques used in big data analytics are more sophisticated than typical business intelligence slicing and dicing (e.g. sales by region or by country). Big data analytic problems typically revolve around sifting through large volumes of data to find things like:

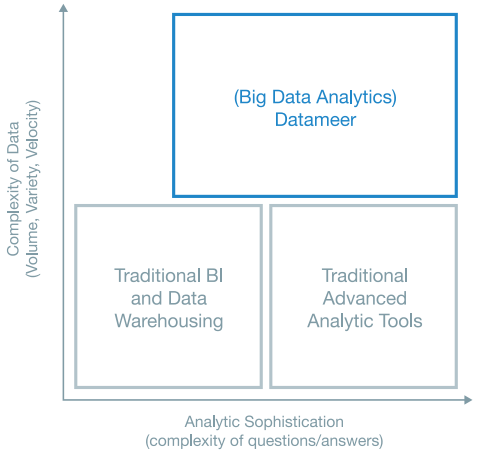
- Hidden patterns (the series of transactions that led to fraud)
- Unknown correlations (what caused someone to buy a product)
- Paths or decisions (what journey someone took to buy a product)
- Complex trend and flow curves (game flow within an online game)
- Clusters of attributes (customer segmentation)

Sophistication of analysis also includes being able to search for the unknown. Analysts don't always know how to answer the analytic question at hand. In particular, what data or combination of data will provide the right results and outcome?

Data Complexity

Data complexity has two different aspects that combine to make a big data analytic problem: large sizes — terabytes to petabytes — and different types, including structured, semi-structured and unstructured. Combined, both of these aspects complicate how an analytic engine deals with such data.

The following chart uses complexity of data and analytic sophistication as two axes, and shows where big data analytics resides on those two dimensions relative to other software segments. Traditional BI approaches are fast and easy for everyday slicing and dicing of information on small- to medium-size data sets.



Traditional analytic tools can deliver more sophisticated analyses but can't handle the more complex data characteristics of big data analytics.

Big data analytic platforms such as Datameer are specifically designed to tackle problems that combine complex data with sophisticated analysis.

To help further identify where you might need big data analytics, Table 1 lists a number of big data analytic use cases and the analytic and data complexity in solving those problems.

USE CASE	ANALYTIC APPROACHES	COMPLEX DATA SOURCES
Behavioral Analytics	Identifying hidden patterns and clusters of attributes	Clickstream data, product usage logs, transactions
Customer Journey	Identifying paths across unknown event correlation	Clickstream data, web analytics, purchases
Pricing Optimization	Complex elasticity curves and clusters of attributes	Prices, costs, margins, product attributes, purchases
Fraud Detection	Uncovering paths and identifying hidden patterns	Transaction logs, geo-spatial data, event activity
Network Utilization	Complex trend and flow curves with unknown correlations	CDRs, network equipment logs, customer SLAs

TABLE 1:
Big Data Analytic Use Cases

What's Important in a Big Data Analytic Platform?

Data & Modeling Flexibility

Big data analytics requires extreme flexibility, both in how it handles data and how it models analytic problems. A lack of data and modeling flexibility is why traditional data warehouse and BI platforms can't adequately handle big data analytics.

Traditional BI approaches are very good at everyday analyses because the structure is fixed and the system can build indexes that allow users to slice and dice their way through data very quickly. This is good for analysis in which your attributes and business questions are well-known ahead of time. It allows the BI and data warehousing team to build the attributes and related values into the data warehouse and push specific dimensions into the analytic model (typically OLAP).

But with big data, this rigidity creates problems. If an analyst needs to look at a big data problem with new attributes, they have to wait for the IT team to structure and add this data to the warehouse. In addition, as more dimensions are added and data points grow, the indexes in the warehouse and OLAP engine start to grow exponentially and can't keep up with the volume of information.

Often with big data, analysts don't know how to answer their analytic questions ahead of time. They may have some theories where they can start, but may not know exactly what they're looking for. In analytic terms, they don't know the importance of various attributes and the relationship between attributes that will influence the outcome.

To this end, a big data analytic system needs flexible modeling that can move attributes in and out quickly and allow the analyst to rapidly iterate through a number of scenarios. The technical concept that enables the flexible modeling is schema on read.

With schema on read, when the analyst decides what data to use, the big data analytic platform builds the schema on the fly to support this specific analysis without modifying or replacing the underlying data. The contrasts with the schema-on-write concept used in data warehouses where the schema is fixed and data is written into that fixed schema. With schema on read, the analyst doesn't need to wait for anyone to change the data warehouse and add new supporting data.

Analyst-Friendly

Perhaps the most important capability of a big data analytic platform is being analyst-friendly. It must make a business analyst more productive while performing a wide range of analyses.

The first aspect of being analyst-friendly is covering the entire analytic process — data integration, data preparation, analysis and visualization. This means the tool must be designed for self-service — meaning the analyst can execute the entire process on their own. If the analyst needs to rely on other staff to integrate and prepare the data, or needs to export to visualization tools, the process becomes fragmented and is fraught with delays.

In addition, the analytic UI must allow the analyst to quickly switch between various steps, make adjustments and immediately see downstream impacts. This allows the analyst to quickly iterate through various experiments in their analysis to find the right outcomes, speeding the analytic process.

One element to look for in an analyst-friendly platform is an easy, familiar interface. Microsoft Excel is the most popular analysis tool on the market.

A familiar, Excel-like spreadsheet interface in the big data analytic platform makes the analyst comfortable and immediately productive. The tool should also have a large number of embedded analytic functions to support the sophisticated analyses found in big data analytics.

Lastly, the analytic tool set needs to have eye-popping visualizations that can effectively show the results to business users. Visualizing thousands of data points often requires a different approach than pie and bar charts. Big data problems like visualizing a customer's journey in the buying process require a far different visualization approach than bar charts of a product's sales by quarter. Big data visualizations should allow business users to quickly see what is important from the analysis, identify what they need to do and take action.

Scalability

Big data analytics is about analyzing terabytes to petabytes of data. Simply put, building an analytic system that can do this is difficult. It's also prohibitively expensive to do with traditional data warehouse and BI approaches.

An analytic platform needs a powerful storage and compute engine to store data and perform analysis at scale within a reasonable response time. And as new technologies emerge at the infrastructure level, the analytic platform needs to be able to support these innovations as they reach enterprise-class maturity.

Hadoop is the best and most popular foundation for scalable, big data analytic processing. Hadoop should be at the heart of the big data analytic platform you choose. It was designed from the outset for the cost-effective storage and processing of large volumes of data. It has almost unlimited compute scalability and data scalability into the petabytes. Hadoop utilizes clusters of commodity hardware and provides a powerful and cost-effective alternative to traditional data warehouse and BI systems.

Unlike traditional data warehouses that only deal with structured data, Hadoop is also designed for storing and processing both structured and unstructured data.

Pre-built data models aren't required, which enables analytic platforms to implement the flexible schema on-read capabilities previously mentioned.

An analytic platform should apply the power of Hadoop to all phases of the cycle. Loading petabytes of data and then integrating it requires the substantial horsepower that Hadoop offers. Preparing and analyzing the data with sophisticated analytic functions also requires the powerful, scalable engine that Hadoop provides.

Big data analytic processing is not one size fits all. Different types of analysis will place varying workloads on the underlying analytic platform. And different execution engines such as MapReduce, Tez and Spark are optimized for different analytic workloads. Simply supporting Hadoop and one execution engine is not enough for an enterprise with a multitude of analytic needs.

The analytic platform needs to support multiple execution engines, each designed to effectively process workloads from different analytic jobs. More importantly, the analytic platform should provide a cost-based optimizer that shields the analyst from the nuances of those execution engines and selects the right one for the job as an analytic process unfolds. This makes the job of the analyst easier and eliminates the need for IT resources that tune the jobs for specific engines.

The analytic platform should also future-proof customers when new technologies and execution engines emerge. Since Hadoop and related big data infrastructure projects are constantly evolving, new execution engines are constantly emerging that help solve different types of analytic jobs. The analytic platform architecture should allow the easy incorporation of these new engines as they mature, and do so in a manner that's transparent to the user and non-disruptive to the applications. Enterprises can then receive new technology in the platform with full protection of their existing investment.

Data Diversity

A common characteristic of big data analytics is data in a variety of diverse formats — structured, semi-structured and unstructured. A big data analytics platform based on Hadoop offers a tremendous advantage as Hadoop provides a complete foundation for storing and processing these diverse data sets.

While Hadoop requires no pre-built data models, it only provides storage and compute capabilities. It lacks data loading and integration tools for end users. This is where an analytic platform adds value.

An analytic platform needs to offer a complete set of data-loading functionality and pre-built connectors for structured and unstructured data that span the many data sources companies use today. To empower and enable less technical users, the platform should also offer a wizard-based integration approach, in which users simply have to specify where the data is, what data needs to be ingested and on what schedule.

Data used in big data analytics can often be inconsistent, variable and have missing data. The best person to understand the gaps in data is not the IT staffer, but the analyst who knows the meaning of the data and how to fix the problems. To this end, the big data analytic platform needs to allow self-service data ingestion and integration without requiring IT staff to build ETL jobs or write MapReduce code, and have the ability to work with dirty data and allow the analyst to easily profile the data, clean and prepare it for analysis.

In addition to working with various data types, the analytic platform needs to support a number of key aspects to manage the process of working with data, including:

- Placing no limits on the size of data ingested for analysis
- Supporting flexible data retention and management policies
- Providing exception reporting on data ingestion processes

Enterprise-Ready

A final but important aspect of a big data platform is being enterprise-ready. As the analytics get operationalized within the organization, the IT organization will look for an analytic platform to support a number of key features that make it enterprise-ready.

The first aspect of being enterprise-ready is being easy to administer, yet offering all of the key controls that IT will want. The platform needs to offer centralized management and control through a web-based interface. There also needs to be both ad-hoc (interactive) and scheduled jobs for both importing data and executing analysis.

With many companies using big data analytics to support regulatory reporting processes or use personally identifiable information in the analytics, governance is an essential part of the big data analytic platform. Governance is more than just security.

There are five key capabilities to provide a complete set of governance features:

- **Quality and consistency** — We discussed the need for data quality in the analytic process but in certain areas like regulatory reporting, a full assessment of the quality and consistency of data is necessary. Complete data profiling, data statistics and metadata management tools aid this process.
- **Data policies and standards** — IT has a goal to implement data access policies that allow them to manage risk appropriately, while still meeting business needs. They need to set policies that expose a subset of data to specific users and apply masking, anonymization, or aggregation to sensitive data fields. They also need to apply further policies and role-based security at every stage of the pipeline, from integration to export.
- **Security and privacy** — Fine-grained access control is important, both at the row and column level, and with the metadata. Integration with enterprise identity management systems like Active Directory and LDAP is essential, as are role-based controls on downloading or exporting data and accessing administrative functions.

- **Regulatory compliance** — To support regulatory processes, the big data analytic platform needs to support cross-artifact lineage features to facilitate verification of where the data came from and how the results were generated. This can then be combined with all relevant user and system events to produce a full audit trail for regulatory analytics.
- **Retention and archiving** — The analytic platform must support flexible, configurable rules for each imported data set's retention policy. Policies such as keeping data permanently, purging records older than a specific time window, or ones based on the number of ingests or analytic executions should be supported. Additional security rules should allow retired data to be either instantly removed, retained until a specified time or manually removed after system administrator approval.

IT teams will also want to examine the architecture of the platform and the performance implications. And lastly, IT will want a platform that can offer an extensible system with open APIs supporting a variety of languages, and an SDK to add custom analytic functions.

Modern BI Evaluation Criteria Checklist

Modern BI platforms are designed to help you answer a deeper range of questions about your business using more data and a greater variety of data. The objective is to move from an IT-lead approach to an IT-enabled one in which citizen data scientists can easily apply their knowledge of data, analytics and the business to answer critical new business questions.

To achieve this, the modern BI platform you choose should contain key functionality that will allow you to:

- Use more of your valuable data, regardless of size, location and format
- Allow BI and analyst teams to be more productive through an analyst-friendly interface, agile analysis and data discovery
- Quickly tell stories about the data to ensure the results match business needs
- Put your insights to work across the organization through easy operationalization of data, strong governance and industrial strength enterprise features

Read on for specific functional items to look for when evaluating modern BI platforms.

Works With All & Any of Your Data

- ☐ Works with structured, semi-structured and unstructured data
- ☐ Offers a complete set of data loading features designed for end users
- ☐ Has 70+ pre-built connectors for structured and unstructured data that span the many data sources a business will use
- ☐ Has built-in data links to existing Hadoop data
- ☐ Possess a fully self-service data ingestion and integration process — no IT needed for ETL
- ☐ Contains no limits to the size of data ingested
- ☐ Has exception reporting on data ingestion processes

Analyst-Friendly

- ☐ Has a fully self-service interface that enables your citizen data scientists to perform all functions, eliminating dependencies on IT
- ☐ Offers a familiar Excel-like spreadsheet-style interface
- ☐ Allows the analyst to apply functions via easy drag-and-drop and point-and-click operations
- ☐ Supports a large number of pre-built analytic functions (270+)

Agile Analysis & Data Discovery

- ☐ Covers the full end-to-end analytic workflow: data connectivity, integration, preparation, analysis, visualization and operationalization
- ☐ Has a fluid workflow with the ability to operate on different analytic steps in parallel and immediately see downstream effects
- ☐ Uses dynamic modeling with schema on read
- ☐ Offers an easy way to profile data sets, then clean and prepare the data for analysis
- ☐ Has easy-to-use advanced analytics for time series, graph, path, text and sentiment analysis
- ☐ Possesses smart data discovery via built-in advanced algorithms for clustering, decision trees, data dependencies and recommendations
- ☐ Supports collaboration and sharing analytic datasets with analyst colleagues
- ☐ Works with existing enterprise data warehouses and BI tools to extend existing analytics by adding new data or new analytic operations to the data
- ☐ Can extend an enterprise data warehouse by pushing results to the EDW for centralized management of analytic result sets

Storytelling

- ☐ Contains easy-to-use, built-in linked visualizations
- ☐ Supports over 30 drag-and-drop visualization widgets
- ☐ Has integrated support for geo-spatial data display
- ☐ Automatically changes chart attributes depending on the data
- ☐ Enables free distribution of storyboards to business end users without requiring the user to have a license
- ☐ Contains the ability to add external elements to the storyboard (images, video, etc.)
- ☐ Offers drill down from the infographic into the analytic application data
- ☐ Allows the analyst to create slideshow-style storyboards
- ☐ Contains full HTML5 support for widgets and infographics for responsive design and seamless cross platform support — desktop, tablet and mobile
- ☐ Easy export and integration with leading visualization tools such as Tableau and PowerBI

Data Operationalization

- ☐ Can export data to external systems or formats that external systems can easily import for fast operationalization of insights
- ☐ Offers an easy-to-use REST API for rapid integration with downstream applications and processes to deliver analytic results
- ☐ Has centralized management and control through a web-based interface
- ☐ Supports both ad-hoc and scheduled jobs for importing data, executing analysis, and exporting downstream applications and systems

Governance

- ☐ Fully supports the five pillars to data governance: quality and consistency, data policies and standards, security and privacy, regulatory compliance and retention and archiving.
- ☐ Offers complete data profiling, data statistics and metadata management tools for data quality and consistency
- ☐ Supports data policies that expose a subset of data to specific users and can apply masking, anonymization or aggregation to sensitive data fields
- ☐ Allows policies and role-based security at every stage of the pipeline, from ingest to export
- ☐ Offers role-based controls on downloading or exporting data and accessing administrative functions
- ☐ Supports cross-artifact lineage features that can be combined with all relevant user and system events to produce a full audit trail for regulatory analytics

- ☐ Allows retention policies such as keeping data permanently, purging records older than a specific time window or ones based on the number of ingests or analytic executions
- ☐ Supports security rules for retired data to be either instantly removed, retained until a specified time or manually removed after system administrator approval

Enterprise-Ready

- ☐ Built to easily scale via native integration with Hadoop storage and compute (doesn't just connect to Hadoop)
- ☐ The platform provides linear scalability on large Hadoop clusters of up to 6000 nodes
- ☐ Uses Hadoop parallelism and scaling in all steps of the analytic cycle: data ingestion, data integration, data preparation, analysis and visualization
- ☐ Has an intelligent execution framework that:
  - ☐ Hides the complexity of the underlying technology from the user
  - ☐ Automatically selects the best execution engines based on the characteristics of the analytic workload
  - ☐ Can break jobs down into smaller tasks and execute the individual tasks on the best engine for that task
- ☐ Has a future-ready architecture that can support new execution engines and technology as they mature
- ☐ Can integrate with enterprise identity management systems like Active Directory and LDAP
- ☐ Offers an extensible system with open APIs supporting a variety of languages
- ☐ Contains a plug-in SDK to add custom analytic functions.
- ☐ Offers flexible deployment options:
  - ☐ On-premise, supporting a variety of different Hadoop distributions (Cloudera, Hortonworks, IBM BigInsights, MapR, Microsoft HD Insights, Pivotal)
  - ☐ In the cloud on leading cloud infrastructure such as Microsoft Azure
  - ☐ As a fully managed Software-as-a-Service (SaaS) solution requiring no infrastructure, IT or software management

FREE TRIAL
datameer.com/free-trial

TWITTER
@Datameer

LINKEDIN
linkedin.com/company/datameer

SAN FRANCISCO
1550 Bryant Street, Suite 490 • San Francisco, CA 94103 USA • Tel: +1 415 817 9558 • Fax: +1 415 814 1243

NEW YORK
9 East 19th Street, 5th floor • New York, NY 10003 USA • Tel: +1 646 586 5526

HALLE
Datameer GmbH • Magdeburger Straße 23 • 06112 Halle (Saale), Germany • Tel: +49 345 2795030

SINGAPORE
Datameer Singapore Pte Ltd • 03-20 Galaxis 1, Fusionopolis Place • Singapore 138522, Singapore • Tel: +65 6809 1157

HONG KONG
12/F International Commerce Centre, 1 Austin Road, West Kowloon, Hong Kong • Tel: +852 2824 8646