BUSINESS INTELLIGENCE

A reference architecture for next-generation big data and analytics

For organizations looking to transition from their current data architecture to one that embraces the power and opportunity in big data, this paper explains:

- The current state and value of Hadoop in the enterprise data landscape
- The architectural challenges of using Hadoop for broad-based analytics and what is needed to overcome them
- The benefits of a next-generation data architecture and how to integrate Hadoop
- The case for existing data warehouses and how they fit into the next-generation data architecture

Executive summary

Hadoop is great for data storage and processing. However, the same architectural features that make it great for running complex transformations on massive amounts of unstructured data make it less adequate for interactive business analysis. Hadoop's challenges in serving large numbers of users at business speed and its lack of a business user interface necessitate a new kind of data architecture. Forward-thinking companies that are adopting Hadoop in their next-generation architecture need to adopt a business intelligence (BI) architecture that separates data storage and processing workloads from end-user analytical workloads, giving business users the best of both worlds.

This new architecture must create a user-oriented analytical data tier on top of Hadoop to aggregate and analytically refine data for browsing, exploration, and delivery throughout an organization. The User Data Tier from Birst, an Infor® company, follows this reference architecture, but Birst® goes a step further, tying Hadoop together with the user-oriented data tier. As a result, users can experience both sophisticated, interactive analytics as well as directly connect to Hadoop to access data at its deepest level of detail. Together with its Networked BI and an Adaptive User Experience, Birst's User Data Tier offers a fast, responsive, and high-concurrency analytic environment, empowering business users to interact with Hadoop in business terms they understand with rapid speed.

> “For three successive years, ‘in production’ as a percentage of Hadoop and big data projects has been below 20% in Gartner surveys.”
>
> **Nick Heudecker**
> Hadoop and Spark: Understanding Open-Source Opportunities and Risks, Gartner, Oct 2018

The current state of Hadoop and its data analysis abilities

Across every industry, organizations are focused on putting data at the center of business transformation to better understand their customers, create product and service differentiation, or to simply lower their costs. In this crowded world of information, Hadoop has become the center of data gravity. Whether used for data archiving, augmenting and complementing existing data warehouses, or for more innovative use cases such as machine learning and Internet of Things (IoT) applications, Hadoop has become a core part of modern data architectures.

Despite its power, Hadoop has remained a tool for data scientist and developers and has failed to become widely adopted across the business. According to Gartner, “Hadoop is firmly entrenched in the landscape—it is deployed in thousands of organizations—but using it effectively has proved to be a key challenge, and it is not being deployed with widespread success in many organizations.” [1]

This low rate of success is mainly due to three factors:

1. Hadoop is not designed to answer analytics questions at business speed

As much as Hadoop has provided a highly scalable and cost-effective data storage and processing engine, its data structure is not built for interactive analysis. Most SQL-on-Hadoop engines are designed to do full table scans but tend to be too slow for individual record lookups, range scans, and analytic scenarios. For example, to see weekly close-of-business sales as compared to the month prior, and for products that fall into category x, but don't belong to region r, you need data to be organized in an analytic-ready format. Otherwise, you will be joining many tables and waiting a long time to get your results; and if you change category x to category y, you must rinse, repeat, and start all over again.

2. Hadoop is not built to handle high-volume user concurrency

Today's data-driven businesses require a system that can handle highly concurrent requests, while exhibiting low latency. Hadoop, on the other hand, tends to focus on high throughput; queries can be very complex and touch most, if not all of the data in the system at any time. This is powerful but does not support many simultaneous queries at once. Hadoop's low concurrency results in an inability to power real-time apps, handle many users at once, or deliver reports as updates happen. It isn't suitable for organizations with large user populations who demand information regularly and simultaneously.

3. Hadoop is not consumable for business users

Today, the simplest reporting interface for getting data out of Hadoop is SQL. This is not a skill that business users have. A business user thinks in business terms, for example, in terms of revenue, sales, and customer, and not in terms of joins, WHERE clauses and SELECT statements. As a result, when Hadoop is in play, IT becomes the gatekeeper, increasingly burdened to prepare data for each new business question or use case. This is acceptable for pilot projects, data science, and hypothesis testing, but not suitable for scenarios where insights are needed at the point of impact, every day, where new questions are asked every hour.

To date, even though some BI and analytics platforms have tapped into Hadoop data lakes to make Hadoop data more impactful, these efforts have either provided a sub-optimal analytic environment with slow response times or have required development teams to painstakingly become the reporting factory that supports business users' decision-making on a daily basis.

This is not sustainable. Organizations realize that their existing data systems are not sufficient to keep pace with their information needs, and therefore are turning to a new approach that delivers value to all users, at scale.

Next-generation data architecture and the user data tier

To move into the next generation of big data platforms that deliver value to a broader organization, Birst has designed a User Data Tier on top of Hadoop and other data sources.

Birst Automated Data Refinement (ADR)

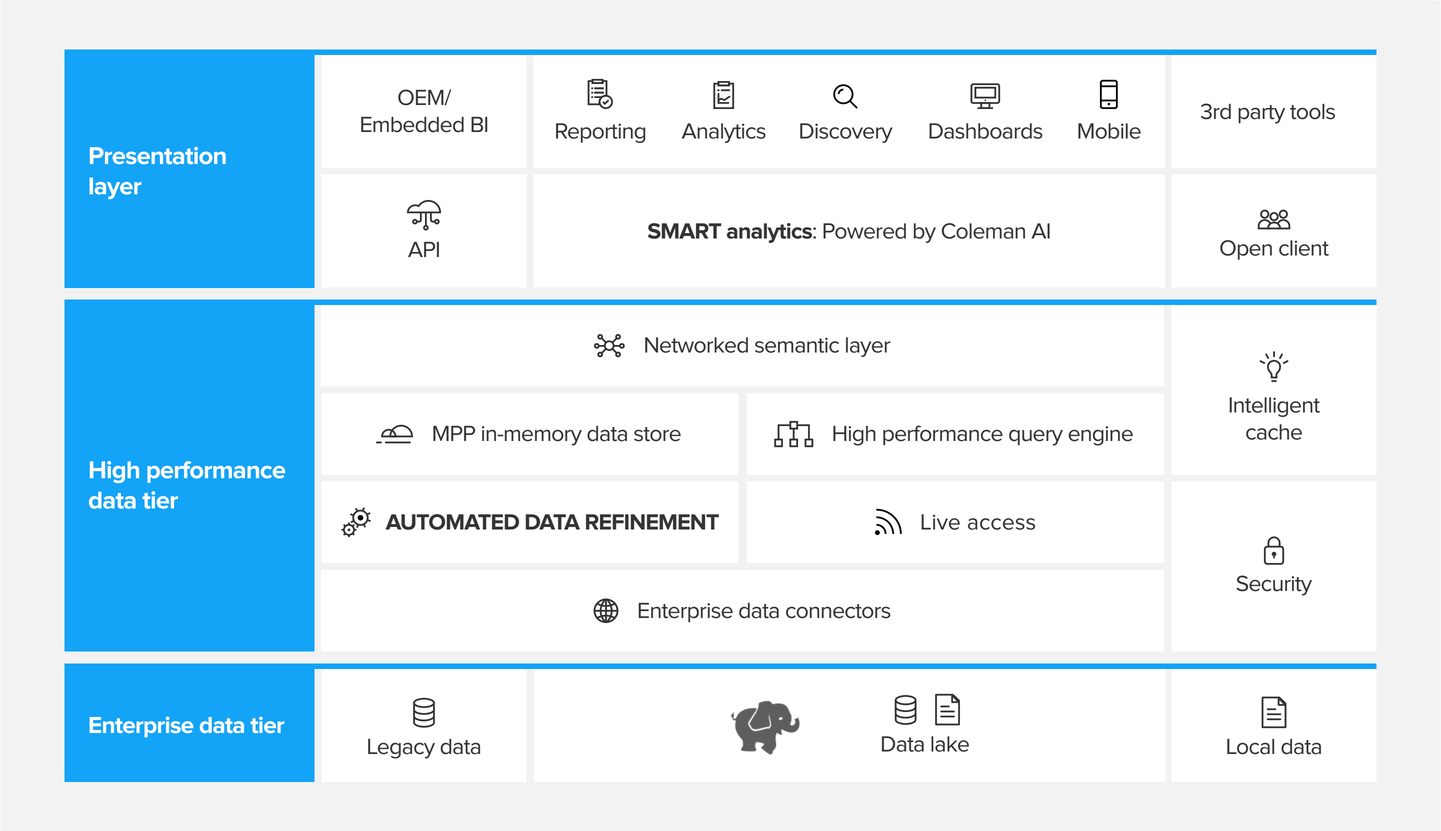
Birst's patented Automated Data Refinement (ADR) technology automatically refines raw data into a user-ready format, builds data hierarchies (e.g., product categories to products to individual SKUs), time-series measures (e.g., trailing three, six, or twelve months), and historical snapshots (e.g., today's data vs a year ago, or a month ago). Birst ADR can build this analytic-ready data model from one source or multiple sources. If multiple sources are at play, Birst ADR combines and organizes data from disparate sources into a single analytic format, conforming and relating information together automatically.

It also creates incremental loads as changes happen in the underlying data sources, reducing data load and processing times.

This automation provides a fast and accelerated approach for creating a user-ready data tier on top of a Hadoop infrastructure, generating aggregated summaries, business dimensions, and measures that enable rich interactive analysis on top of data in Hadoop and other sources.

Unlike Hadoop that has high-latency and low-concurrency, this analytic data store offers interactive and fast query response times, can analyze large data sets quickly, and handles "what if" scenarios easily. If an individual analyst or business user veers away from searching the aggregated, hierarchical data available in the analytic data store, Birst can automatically route the query back to a low-cost data store like Amazon Redshift or even all the way back to the raw Hadoop data source for very detailed analysis. This query navigation process is seamless to the user and eliminates the need to drop into a primitive and unfamiliar set of Hadoop frameworks.

Birst High Performance Big Data Edition



Presentation layer

- OEM/ Embedded BI
- Reporting
- Analytics
- Discovery
- Dashboards
- Mobile
- 3rd party tools
- API
- **SMART analytics**: Powered by Coleman AI
- Open client

High performance data tier

- Networked semantic layer
- MPP in-memory data store
- High performance query engine
- Intelligent cache
- **AUTOMATED DATA REFINEMENT**
- Live access
- Security
- Enterprise data connectors

Enterprise data tier

- Legacy data
- Data lake
- Local data

Birst intelligent multi-layer cache, aggregates, and query federation

Birst's User Data Tier interprets the best execution path for queries and decides the most cost effective and performant path to get answers to business users. Birst employs caching and aggregate awareness to send queries to the cache first, then data in the User-Ready Data Store. If data is not cached, Birst generates one or more queries depending on how data is sourced. For example, for federated queries where some data is stored in the Birst User-Ready Data Store and some in external sources, Birst automatically sends a query to both and combines the results for the business user.

Birst's in-memory caching includes both "exact" and "fuzzy" matching. Exact matching handles situations where queries match a prior query exactly, and fuzzy match covers scenarios where the query is a subset of prior queries.

Birst Live Access

In cases where the scenario requires data directly from Hadoop, Birst Live Access allows a direct connection to Hadoop using popular SQL-on-Hadoop frameworks. This is most useful when Birst is used for data at higher granularity levels (e.g., day, week, month, quarter, and year) and data at lower granularity levels (e.g., intraday transactions) is kept in Hadoop. Birst's Query Navigator will select the higher granularity data from the Birst analytics data store but will use Birst Live Access to get the detailed data directly from Hadoop in real time.

Birst Semantic Layer

To hide the complexities in data, expose data in simple business terms, and enable business users to create their own metrics on the fly, Birst puts a semantic overlay on top of its User Data Tier. Since non-technical users typically lack SQL skills, they are not able to query Hadoop directly. A semantic layer enables those users to think in business terms, instead of learning SQL-on-Hadoop skills. For example, a semantic layer may contain information about high-value customers. This information could have gone through different levels of calculation and transformation—all seamless to the business user who simply searches for the words "customer" or "high-value customer" and gets accurate results.

A semantic layer abstracts the users' interaction with data from the data itself. For example, imagine that your historical purchase data is stored in a data warehouse, current sales data is in a customer relationship management (CRM) application, and website clickstream data is in Hadoop. A semantic layer allows a business user to ask for today's online purchases from recurring customers in the Northeast by simply dragging and dropping information about "online purchases" and "recurring customers" and then filtering by "date" and "region." The alternative, and without having a semantic layer, the user would need to compose three separate queries, get the results back, and rationalize what records are actually from recurring customers, what purchases are coming from the online store, and conform the dimensions manually. Simple? Hardly!

Birst Networked BI

For end users and decentralized groups to take advantage of the data in Hadoop, Birst, as explained in the previous section, transforms Hadoop's raw data into a User-Ready Data Store and exposes it to end users through a semantic layer. However, to allow business teams to work on their own, while staying networked to a central, governed data set, Birst takes this a step further by providing Networked BI.

In this model, different groups, such as finance, customer support, sales, and marketing use their virtual copies of the User Data Tier to gain access to centralized data (e.g., historical or corporate data, stored in Hadoop) and blend that with their local data and spreadsheets. Since these virtual instances are logical, and not physical, there is no need to recreate the analytic environment. It is as simple as clicking a button and letting individual business groups explore information on their own. This paradigm creates consistency and collaboration between IT and business teams, ensures centralized governance, and empowers data ownership, independence, and self-service data blending at the point of impact.

Birst Adaptive User Experience

Birst delivers an Adaptive User Experience that uniquely meets the individual needs of all users by supporting different styles of analytics, including pixel-perfect reports, highly interactive and responsive dashboards, intuitive visual discovery, native and offline mobile access, embedded analytics, predictive analytics, and data mashups. Each style is not a separate tool. Birst blurs the lines between these interfaces, enabling users to simply interact with data as they move from discovery to dashboards to reports, creating, collaborating, and publishing with a single click. Since each of these styles pulls data directly from the Birst Semantic Layer, while the data presentations vary from one interface to the other, data values remain consistent across all form-factors.

Birst also provides an Open Client Interface that allows companies that utilize other data visualization and discovery clients like Excel to access Birst through its Semantic Layer. This ensures data consistency and trust among different users, regardless of what tool they choose to interact with the data.

Using Birst Adaptive UX, all users, whether individual analysts, line of business teams, executives, and even customers can take advantage of data in Hadoop, without having to learn new skills.

What happens to existing data warehouses?

The short answer: You can decide what to do later.

Many enterprises have made significant investments in data warehouse infrastructures. Disregarding these investments to switch to an all-Hadoop infrastructure is more than what many are willing to do today. So even though Hadoop is rapidly maturing, and its use cases are proliferating and providing more value, some organizations feel the need to plan carefully for the future and select an analytic platform that gives them the flexibility to work with their existing data warehouse, as well as emerging systems like Hadoop.

Birst gives you the option to do just that. You can connect to an already pre-built data warehouse in real time via Birst Live Access, without having to reinvent the wheel or recreate another analytic-ready data store. You can use Birst query federation capabilities to combine data from Hadoop and legacy warehouses. This helps you extend the investment in your existing data warehouses, while also starting to reap the benefits of emerging technologies such as Hadoop.

Vertafore

Vertafore turns raw data into an analytic product

Vertafore, the leading provider of insurance software and technology, connects independent brokers with carriers at every point of the distribution channel.

Vertafore's unique footprint in the industry provides the company with data from 20,000+ agencies, 1,600 carriers, and 230+ million policies. The goal was to leverage all this data into an analytics product that would scale to thousands of clients. With Birst, Vertafore Analytics easily delivers big data into the hands of insurance carriers and independent agencies—giving them unprecedented control.

Closing thoughts

Next-generation data architectures allow you to separate data storage and processing workloads from end-user analytical workloads, creating the optimal environment where Hadoop is leveraged to its full potential. Birst's User Data Tier, together with its Networked BI and Adaptive User Experience offers a fast, responsive, and high-concurrency analytic environment on top of Hadoop, empowering business users to interact with Hadoop in business terms they understand with rapid speed.

Learn more ›

References

[1] Merv Adrian, Arun Chandrasekaran, Adam Ronthal, FAQs on the Future of Hadoop, Gartner, Oct 2018.

birst
an Infor® company

Infor builds business software for specific industries in the cloud. With 17,000 employees and over 68,000 customers in more than 170 countries, Infor software is designed for progress. To learn more, please visit www.infor.com.

Follow us :

641 Avenue of the Americas, New York, NY 10011