Tableau and Big Data: An Overview

Table of Contents

What big data looks like today

The evolution of data and demand for analysis

Data is everywhere—so is the demand to access and analyze it. "Big data" as a buzzword may have settled down, but the "three Vs" of big data—volume, variety, and velocity—apply more than ever to big data analytics use cases. Though subjective, these and other Vs the industry has discussed (like variability, validity, veracity, etc.) serve to remind us that big data today is still simply data—it's just gotten so complex that organizations must innovate to effectively gather, curate, understand, and make use of it.

Digital transformation is happening across every industry and all sizes of organizations with a multitude of "things" creating massive amounts of data in many formats and sources. Organizations are collecting, processing, and analyzing more diverse data than ever before. From schema-free JSON to nested types in other databases like relational and NoSQL, to non-flat data—like Avro, Parquet, XML, etc.—data formats are multiplying and connectors are crucial to make use of them.

Organizations often have a combination of the following:

- **Structured data** with precomputed aggregates to specific questions, perhaps pulled as extracts for in-memory computing, and aggregated for analysis. This is typically the most refined and easily accessible data an organization has.
- **Semi-structured data** (or object storage) perhaps in relational databases, data warehouses, or data marts. Often, these are regularly refreshed business concepts for entity analysis—known questions with unknown answers—for example, transactions, opportunities, or actions taken by individual salespeople on opportunities.
- **Raw, unstructured data** in a data lake or cloud storage. This includes stream data created by social network feeds, IoT devices, and more. Data scientists may mine and transform this data, but its full potential is still unknown.

While some data has yet to find its most valuable use cases, all of this data is met with a greater demand for knowledge workers to access and analyze it for decision-making. The applications used for data analysis and visualization are gravitating toward the data itself. This means a large-scale shift towards the cloud, where analysis can occur alongside robust storage and data processing services that allow for greater flexibility and scale. Whether an organization has an extensive, cloud-based big data practice or is currently doing very little analysis of their data, they can reap significant benefits by giving people across business and IT departments the ability to visualize patterns and analyze for insights it contains.

In spite of modern analytics bringing broader capabilities to more business users of all skill levels, finding ways to make all of this data a useful resource for the entire organization presents many complex challenges. Business needs change as often as the data itself, necessitating a big data strategy and architecture that are agile and adaptable. Rather than building monolithic platforms with a focus on data connectivity, organizations would be wise to widen their scope of the big data opportunity and think about its evolving analytics use cases. Otherwise they risk missing the bigger picture.

Big data is both a promise and a peril

Data assets are increasingly becoming a key area of differentiation between wildly profitable and struggling businesses. However the massive scale, growth and variety of data are simply too much and too expensive for relational database management systems to handle. In addition to hardware cost savings due to precomputation and shared computation, customers also seek to minimize moving their data around. Infrastructure that allows them to move data in the most agile ways will help to address the gap between raw, unstructured data and data that's ready for users to analyze.

Organizations also face issues with connectivity and performance. Even with options for live connections or in-memory analysis, huge data lakes can be heavy on operations to generate extracts, or to blend with other data. A modern and self-service approach to analytics has many promises of agility, but making massive joins on these datasets can choke up the system.

IT and the business must work together, but with a bottom-up methodology comprised of subject matter experts creating metadata, business rules, and reporting models. These processes must constantly iterate and improve to meet the evolving needs of the business; in today's era of digital transformation, the business won't stand still, so your big data analytics framework shouldn't either.

How Tableau works with big data

The big (data) picture

Everything we do at Tableau supports our mission to help people see and understand their data. Tableau is the modern analytics platform for the digital economy because we fundamentally believe in the democratization of data. The people who know the data should be the ones empowered to ask questions of the data, meaning knowledge workers of all skill levels should have the ability to access, analyze, and discover insights of their data wherever it may reside.

As many customers are dealing with a diverse set of big data technologies, we have aligned our engineering investments, partnerships within the ecosystem, and overall vision with the evolution of the data landscape. Tableau has a rich history of investments ahead of the curve in big data. These investments include data connectivity to both Hadoop and NoSQL platforms, as well as large-scale on-premises and cloud data warehouses.

> We started off with a very narrow business use-case and then it just quickly spread. Tableau makes it simple and simplicity, everyone wants talk about big data analytics but Tableau simplifies it.
>
> —ASHISH BRAGANZA, DIRECTOR OF GLOBAL BUSINESS INTELLIGENCE, LENOVO

Learn how Lenovo increased reporting efficiency by 95% across 28 countries

Data access and connectivity

To enable analysis of data of any size or format, we support broad access to data wherever it lives. Tableau supports over 75 native data connectors today as well as countless others through our extensibility options. As new data sources emerge and become valuable to our users, we continue to integrate and certify vendors' connectors with Tableau, incorporating them into our product to lower the friction for accessing data. We believe there is and always will be many sources of data that one person wishes to use—whether web traffic, records in databases, log files, and so on.

- **SQL-based connections** — Tableau uses SQL to interface with Hadoop, NoSQL databases and Spark. The SQL that Tableau generates is standardized to the ANSI SQL-92 standard. Using SQL is powerful because it is extremely compact (one expression), it is open source and standardized, there are no library dependencies and it is very rich and expressive. For example, using SQL, one can express join operations, functions, criteria, summarization, grouping and nested operations.

- **NoSQL interfaces** — Just as the name implies, NoSQL ("not only SQL") databases can have data that is modeled in nonrelational in addition to relational formats, supporting additional storage types including column, document, key-value, and graph. It also means they can support SQL-like interfaces.

- **ODBC** — Tableau uses drivers leveraging the Open Database Connectivity (ODBC) programming standard as a translation layer between SQL and SQL-like data interfaces provided by these big data platforms. By using ODBC, you can access any data source that supports the SQL standard and implements the ODBC API. For Hadoop, this includes interfaces such as Hive Query Language (HiveQL), Impala SQL, BigSQL and Spark SQL. To achieve the best performance possible, we custom tune the SQL we generate as well as push down aggregations, filters, and other SQL operations to the big data platforms.

- **Web Data Connector** — With the Tableau Web Data Connector SDK, people to build connections to data that lives outside of the existing connectors. Self-service analytics users can augment their big data analysis with outside data by connecting to almost any data accessible over HTTP, including internal web services, JSON data, and REST API.

Fast interaction with all data at scale

We want users to have access to all their data, at scale, to integrate with other data, and find insights fast. To help make self-service, visual analytics possible with big data, Tableau has invested in several pioneering technologies.

- **Hyper data engine** — Hyper is our high-performance in-memory data engine technology that helps customers analyze large or complex data sets faster. With proprietary dynamic code generation and cutting-edge parallelism techniques, Hyper better utilizes modern hardware for up to 3X extract creation and 5X query speed than the previous Tableau Data Engine. Hyper can also augment and accelerate slower data sources by creating an extract of the data and bringing it in-memory.

- **Hybrid data architecture** — Tableau can connect live to data sources or bring data (or a subset) in-memory. You can go back and forth between these modes to suit your needs. Our hybrid approach to accessing data brings a lot of flexibility for users and can help to optimize query performance.

- **VizQL™** — At the heart of Tableau is a proprietary technology that makes interactive data visualization an integral part of understanding data. A traditional analysis tool forces you to analyze data in rows and columns, choose a subset of your data to present, organize that data into a table, then create a chart from that table. VizQL skips those steps and creates a visual representation of your data right away, giving you visual feedback as you analyze. VizQL allows you limitless exploration your data to find the best representation of it—and with unlimited "undo," there is no wrong path. In this cycle of visual analysis, users learn as they go, add more data if needed, and ultimately get deeper insights. It's not only a richer experience, but one more accessible to all skill levels than to build dashboards by code.

> With Tableau, you can actually interact with the data set in real time and you are able to analyze and then present it in the way that you want within a few minutes.
>
> —JAMIE FAN, PRODUCT ANALYTICS LEAD, GRAB

Learn how Grab analyzes millions of rows of data to improve customer experiences

Tableau and the big data analytics ecosystem

A modern analytics platform like Tableau may be the key to unlocking big data's potential through discovering insights, but is still just one of the critical components of a complete big data platform architecture. Putting together an entire big data analytics pipeline can seem like a challenge in itself. The good news is that you don't need to build out the whole ecosystem before you get started, nor do you need to integrate every single component for an entire strategy to get off the ground.

Tableau fits nicely in the big data paradigm because we prioritize flexibility—the ability to move data across platforms, adjust infrastructure on demand, take advantage of new data types, and enable new users and use cases. We believe that deploying a big data analytics solution shouldn't dictate your infrastructure or strategy, but should help you to leverage the investments you've already made, including those with partner technologies within the big data ecosystem.

Cloud infrastructure

Organizations are increasingly moving business processes and infrastructure to the cloud. As cloud-based infrastructure and data services have removed some of the major hurdles faced with on-premises Hadoop data lakes, cloud-based big data analytics solutions are easier to implement and manage than ever before.

> **Hadoop laid the foundation for the modern data lake with its powerful combination of low-cost, scale-out storage (Hadoop Distributed File System—HDFS), purpose-built processing engines (first MapReduce, then over time Hive, Impala, and Spark), and shared data catalog (Hive metastore).**
>
> **Today, the once co-located storage and compute services can scale as needed and independently in the cloud. Resources also scale up and down a lot more easily, and with on-demand pricing. Overall, the cloud makes for greater efficiency, management, and coordination of services.**
>
> Learn more in this great article from Josh Klahr, VP of Product at AtScale.

Tableau delivers key integrations with cloud-based technologies that organizations already use, including Amazon Web Services, Google Cloud Platform and Microsoft Azure.

Ingest and prep

In modern ingest-and-load design patterns, the destination for raw data of any size or shape is often a data lake: a storage repository that holds a vast amount of data in its native format, whether structured, semistructured, or unstructured. Data lakes support modern big data analytical requirements through faster, more flexible data ingestion and storage for anyone to quickly analyze raw data in a variety of ways.

Stream data is generated continuously by connected devices and apps located everywhere, such as social networks, smart meters, home automation, video games, and IoT sensors. Often, this data is collected via pipelines of semi-structured data. While real-time analytics and predictive algorithms can be applied to streams, we typically see stream data routed and stored in raw formats using lambda architecture and into a data lake, such as Hadoop, for analytics usage. Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch and stream processing methods. The design balances latency, throughput, and fault tolerance challenges. A variety of options exist today for streaming data including Amazon Kinesis, Storm, Flume, Kafka, and Informatica Vibe Data Stream.

Data lakes also provide optimized processing mechanisms via APIs or SQL-like languages for transforming raw data with "schema on read" functionality. Once data has landed in a data lake, it needs to be ingested and prepared for analysis. Tableau has partners like Informatica, Alteryx, Trifacta, and Datameer that help with this process and work fluidly with Tableau. Alternately, for self-service data prep, you can use Tableau Prep.

Storage and processing

Hadoop has been used for data lakes due to its resilience and low cost, scale-out data storage, parallel processing, and clustered workload management. While Hadoop is often used as a big data platform, it is not a database. Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, massive processing power, and the ability to handle extreme volumes of concurrent tasks or jobs.

In a modern analytics architecture, Hadoop provides low-cost storage and data archival for offloading old historical data from the data warehouse into online cold stores. It is also used for IoT, data science, and unstructured analytics use cases. Tableau provides direct connectivity to all the major Hadoop distributions with Cloudera via Impala, Hortonworks via Hive, and MapR via Apache Drill.

There will always be a place for databases and data warehouses in modern analytics architecture, and they continue to play a crucial role in delivering governed, accurate, conformed dimensional data across the enterprise for self-service reporting. Even companies who adopt other technologies (e.g. Hadoop, data lakes) typically retain relational databases as a part of their data source mixture. Snowflake is one example of a cloud-native SQL-based enterprise data warehouse with a native Tableau connector.

Object stores, such as Amazon Web Services Simple Storage Service (S3) and NoSQL databases with flexible schemas can also be used as data lakes. Tableau supports Amazon's Athena data service to connect to Amazon S3, and has various tools that enable connectivity to NoSQL databases directly. Examples of NoSQL databases that are often used with Tableau include, but are not limited to, MongoDB, Datastax, and MarkLogic.

The data science and engineering platform Databricks offers data processing on Spark, a popular engine for both batch-oriented and interactive, scale-out data processing. Through a native connector to Spark, you can visualize the results of complex machine learning models from Databricks in Tableau.

Query acceleration

While you can perform machine learning and conduct sentiment analysis on big data, the first question people often ask is: How fast is the interactive SQL? SQL, after all, is the conduit to business users who want to use big data for faster, more repeatable KPI dashboards as well as exploratory analysis.

This need for speed has fueled the adoption of faster databases leveraging in-memory and massive parallel processing (MPP) technology like Exasol and MemSQL, Hadoop-based stores like Kudu, and technologies that enable faster queries with preprocessing like Vertica. Using SQL-on-Hadoop engines like Apache Impala, Hive LLAP, Presto, Phoenix, and Drill, and OLAP-on-Hadoop technologies like AtScale, Jethro Data, and Kyvos Insights, these query accelerators are further blurring the lines between traditional warehouses and the world of big data.

Data catalog

Enterprise data catalogs essentially serve as a business glossary of data sources and common data definitions, allowing users to more easily find the right data for decision making from governed and approved data sources. They are populated with metadata from tables, views, and stored procedures by scanning ingested data sources. Data curation efforts can even go so far as to include knowledge base information and web links to help users understand the context of the data and enable more intelligent classification and automated data discovery.

Data catalogs exist within visual analytics solutions and are also available as standalone offerings designed for seamless integration with Tableau. Some of our data catalog partners include Informatica, Alation, Unifi, Collibra, and Waterline.

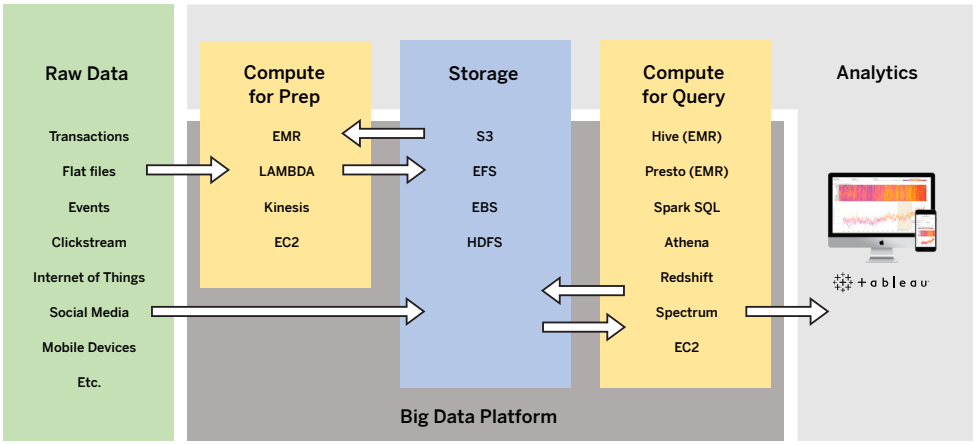
Big data analytics architectures

It's important to remember that there is no "one size fits all" when it comes to successful big data architectures. Our customers have unique, tailored solutions for their big data analytics, with different platforms and tools to make up their data pipelines. That said, we'll follow up with some observations about shared components of the architectures that have contributed to the success of these big data analytics platforms.

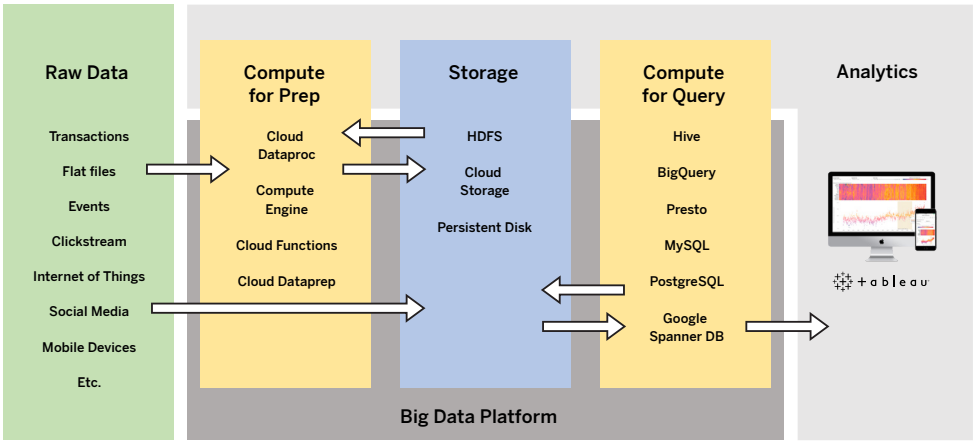
Disclaimer: *Please note the following examples are Tableau interpretations and were not designed by the cloud providers or customers they represent. Where available, we've included links to original illustrations. These diagrams are simplified generalizations meant to highlight similarities in key elements of different flows. They may not reflect every piece of the full big data analytics platform and may only represent certain use cases. Note also that "compute for prep" most resembles "process/catalog" where "compute for query" most resembles "analyze/model."*

Major cloud provider examples

Amazon Web Services



Google Cloud Platform



Microsoft Azure

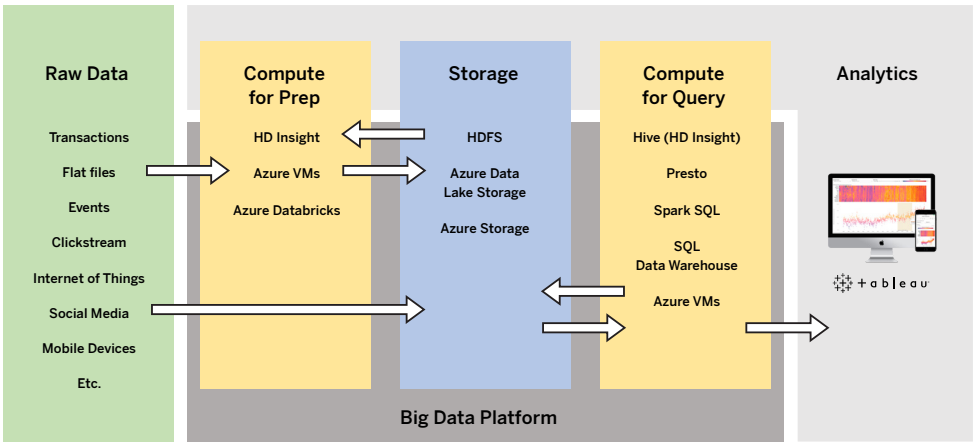
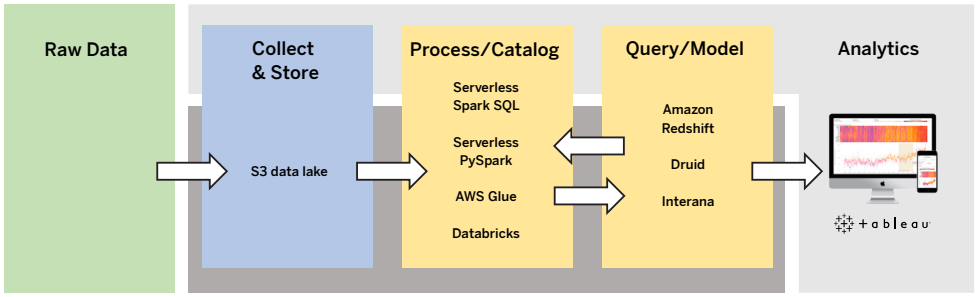
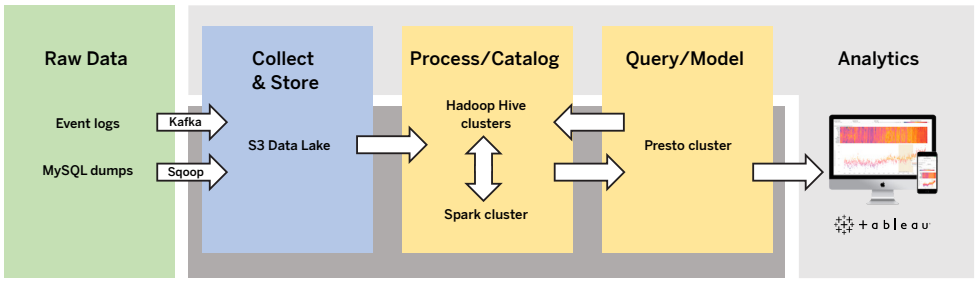


Tableau customer examples

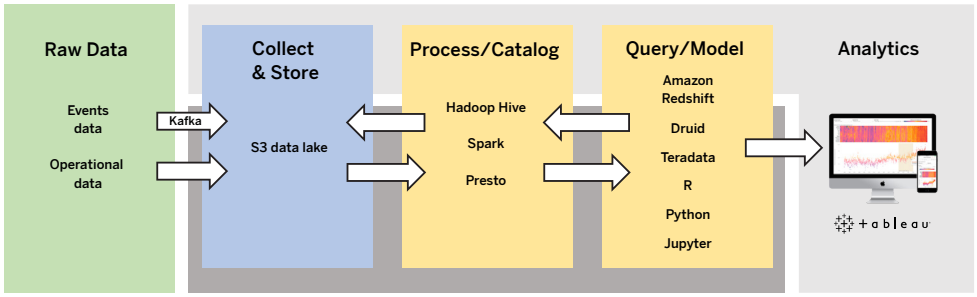
Edmunds – Learn more



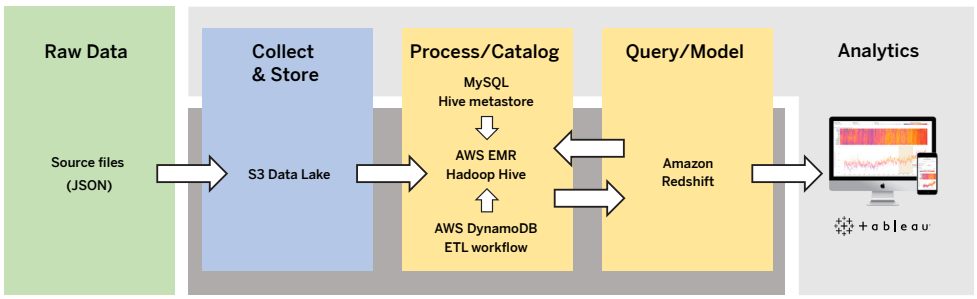
Airbnb – Learn more



Netflix – Learn more



Expedia – Learn more



Common patterns

While no two enterprise architectures are the same, noting similar patterns and what they share in common can help you in strategizing your own big data analytics platform. Here is what we've observed consistently in successful big data analytics architectures:

- **A storage layer** — Many call it a data lake. Your data strategy may necessitate multiple storage environments, but should comprise structured, semi-structured, and unstructured data.
- **Server and serverless compute engines** — Some for data preparation and analytics, other compute engines for querying. The dynamic nature of serverless compute allows for more flexibility and elasticity, as there is no need to pre-allocate resources.
- **Support for volume, velocity and variety** — This applies not just to data, but its growing complexity and number of use cases, some of which are yet to be discovered.
- **The right tool for the job** — It's important to adapt the components of your architecture to address your unique data strategy, but it's also critical to remain agile in the face of changing business needs.
- **Enterprise-level governance and security** — While we haven't gone into much detail in these areas, security and governance are foundational for ensuring scalability and proper use of your data.
- **Cost consciousness**— Take cost into account when considering the necessary power and flexibility for your big data architecture. The cloud affords a lot of elasticity for growth, but you'll want to consider the financial implications of your data storage and processing, concurrency, latency, analytics use cases, etc.

As the big data landscape continues to evolve, one theme is consistent throughout all challenges: Businesses need to be able to use a common, modern analytics platform to access their data, whether big or small, and wherever it lives. With the right platform, processes, and programs to empower people, data-driven decision making will prove a tremendous asset.

About Tableau

Tableau is a complete, easy-to-use, enterprise-ready visual business intelligence platform that helps people see and understand data through rapid-fire, self-service analytics at scale. Whether on-premises or in the cloud, on Windows or Linux, Tableau leverages your existing technology investments and scales with you as your data environment shifts and grows. Unleash the power of your most valuable assets: your data and your people.

Additional resources

Building blocks of a modern analytics platform

Tableau enterprise analytics powered by IT

Tableau for the enterprise: IT overview

Tableau free trial

tableau