cloudera

WHITE PAPER

The Many Advantages of Cloudera Enterprise

The Modern Platform for Machine Learning and Analytics Optimized for Cloud

INSIGHTS

Platform Matters

There are many reasons why your choice of data platform matters. It will serve as the foundation for your digital transformation, enabling you to gain actionable insight and drive immense and measurable value back to the business.

Our definition of data platform is an integrated set of capabilities and functions that drive analytics and data management. This should be oriented around how diverse groups use data to gain insights. A legacy platform (such as a relational database or enterprise data warehouse) is typically too narrow and rigid to meet modern requirements. These requirements include extreme scale, speed of analytics, support for diverse data types, operating environments, and more. A modern approach combines different analytics disciplines into one unified and flexible platform that runs anywhere - on-premises, public cloud, private cloud, or hybrid. This approach makes it easier for BI analysts, data scientists, data engineers, admins, and other knowledge workers to get new insights and innovate in their business.

Your Data-Driven Journey

Most businesses have a variety of initiatives under the broad category of analytics. These can align with different goals. Examples include growing the business with more efficient sales and marketing, connecting products and services as in machine to machine communication and IoT (Internet of things), or protecting the business by reducing risk and preventing cyber attacks.

These organizations start with a goal like achieving greater visibility at a lower cost, which can be accomplished by loading and complementing an existing data warehouse. Then once they have familiarized themselves with the new platform and found success, they will usually expand into many new areas. This journey doesn't end, as you can always find more value and the more things you do, the better your return on investment.

The Pain of Analytics Silos

Here are some of the challenges for different groups within your business, associated with legacy platforms and the silos they inadvertently create.

1. For IT infrastructure and operations teams:
   - Single-use, inflexible data sources such as traditional databases, data warehouses, and data marts are hard to manage individually, much less as a whole.
   - Redundancy and fragmentation of data sets across these disparate sources create operational inefficiencies and drive up costs.
   - Complexity can lead to errors, delays, and a never-ending backlog of requests for data and maintenance.
   - Blown budgets are often the unfortunate result, both in terms of human and technological costs.
2. For data scientists and business analysts:
   - Users often can't find the data they need and end up waiting on IT to fulfill requests, often so long the question is no longer relevant.
   - Too much data preparation effort also delays projects and answers, as each source will need to be normalized and joined before it can be analyzed.
   - Users struggle to with complex queries, as too many tools each only handle a part of the job.

- Time is wasted moving data between analytics environments to serve different stages of the diverse needs.

3. For the head of data and analytics:

- Much of the organizational leaders' time becomes administrative, not innovative in nature. They now risk being outclassed by cloud and the allure of instant resources, even if it doesn't provide a cohesive approach.
- Failing to meet all enterprise business requirements leads to disappointment and stalled initiatives.
- Ultimately, they are likely to be replaced and the organization will start over again.

The benefits of a unified platform

A common frustration is that single use, niche databases and tools are limited in function, allowing only one approach in analytics. This leads to great efforts in trying to integrate disparate products to handle different functions like SQL-driven data warehousing and business intelligence (BI), real-time analytics for IoT, data science machine learning (ML), and/or the data engineering and preparation that precedes each of the above. This need to integrate increases your teams' work and delays projects significantly. Further, product and operating costs rise with each additional vendor offering that must be included. Not least, the risk of errors or security gaps is compounded by multiple distinct security frameworks, each with its own set of capabilities and foibles.

To really understand and predict behavior with analytics you need a more omniscient view, which can only be brought by a unified platform. A unified platform will benefit everyone who analyzes and manages data, making them more productive and in turn drive faster and more differentiated innovation for your business.

Multi-disciplinary Analytics Add More Value

Those same teams could be in a better world, which a modern data platform can help them realize. Cloudera solves those challenges, putting everyone on a path to success and each group of stakeholders will benefit as shown here.

1. For IT infrastructure and operations teams:

- A unified platform will give them time to focus on curating data not firefighting requests.
- Central control and security will provide operational efficiency and significantly reduce risks associated with data breaches or failed compliance audits.

2. For analytics users such as data scientists and business analysts:

- They can find valuable insights faster from tapping one source of truth, not dragging data from many silos.
- They can bring the best tools per job, and utilize a wide range of analytics approaches in concert to solve bigger questions than any one method alone.

3. For the head of data and analytics:

- A modern platform will help them drive the business to be agile and efficient in gaining new insights.
- They can enable teams, advise on analytics strategy, and productionalize innovations as new applications creating differentiation for the business.

Three Macro Trends Changing The Game

Machine Learning

One major focus for Cloudera Enterprise is machine learning. Our data platform helps you model your business and make predictions that traditional approaches can't handle. It gives your data scientists access to new data sources they can combine to find new patterns, without increasing risk or liability. The platform also enables them to work directly with ALL data, not a sample downloaded to a laptop. Not least, the complete data pipeline can be built and managed in one platform, not shuttling data around a bunch of silos.

Analytics

Another focus is in enhancing data warehouses with a more modern solution. Traditional enterprise data warehouses are rigid, costly, limited in size by that cost and architecture, and limited in scope to well-defined structured data. They still have a place though, and Cloudera Enterprise complements them, especially for exploration and self-service analytics. Cloudera Enterprise can also work with less structured data and combine approaches, not just SQL, but also additional techniques like integrated search and data science.

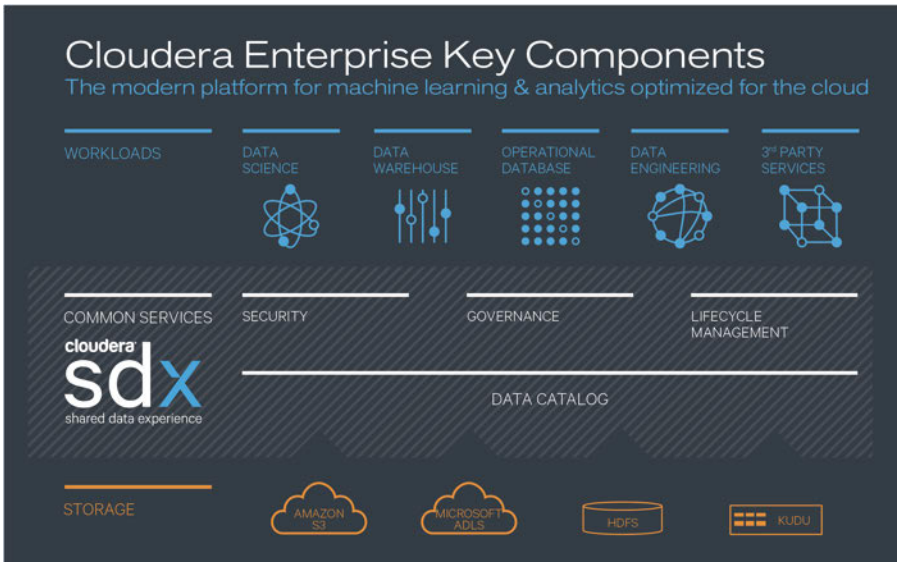
Cloud

A third area is the recognition of how cloud promises to change the game. Unfortunately, many cloud analytics services aren't as elastic as you'd think. Try to shrink AWS Redshift or grow to petabytes for example, and you will find the limitations. Cloudera Enterprise fixes those limitations with an enterprise-grade offering that leverage cloud infrastructure, AND delivers on the promises. The platform also lets you run the same solution everywhere, across hybrid- and multi-cloud environments, for operating efficiency and to avoid cloud vendor lock-in. One step further towards simplifying is to automate much of the management, as we do with Cloudera Altus.

How Cloudera Enterprise Delivers Value

The diagram below shows a functional view of what we offer with Cloudera Enterprise. Note that 3rd party partners can operate and enhance capabilities at all levels. Our platform works on your choice of storage, on-premises or cloud, and supports all of the most common analytics disciplines. These approaches include Data Science (for ML), Data Warehouse (for data warehouse and BI), Operational DB (for real-time streaming), data engineering (for ETL and more), and integrated search.

The Cloudera SDX layer (Shared Data Experience) makes it possible for companies to run dozens - hundreds - of these multi-disciplinary analytics workloads against a common pool of data. SDX applies a centralized, consistent framework for catalog, security, governance, data lifecycle and more. SDX makes it faster, easier, and safer for organizations, teams, people to develop and deploy high-value, multi-function use cases.



Analytics Workloads on Cloudera Enterprise

Data Science

Cloudera's unified platform for data and machine learning eliminates silos and speeds time to value -- on premises or in the cloud. Having secure, self-service access to enterprise data makes data scientists more productive without introducing risk. Access to elastic on-demand provisioning of resources delivers the computing power data scientists need for the most demanding analysis. A free choice of Python, R, and Scala support brings the flexibility, innovation, and value of open source machine learning tools and libraries to the data. Not least, containerized environments simplify collaboration, sharing, reproducibility, dependency and deployment management. Altogether, the platform helps you accelerate data science from research to production.

Data Warehouse

Cloudera Data Warehouse delivers high-performance SQL for ad hoc exploration and selfservice BI. It brings flexibility to iterate with more data and more use cases than a traditional data warehouse. Users can also go beyond SQL for shared data with open standard tools, like integrated text search. Data warehouse optimization of your existing databases is facilitated by Cloudera Navigator Optimizer, which can assess individual jobs for ease of migration. Again, unlike a traditional hardware-based data warehouse appliance, with Cloudera Enterprise you can cost-effectively scale both on-premises and leverage transient or persistent resources in the cloud.

Operational Database

Cloudera Enterprise can achieve real-time data serving at scale, injecting the latest information into processes and decisions that are data-dependent. Fast analytics on fast data through a simplified architecture provides the ability to examine trends that include even the most recent data. Analytics application developers can leverage any data source and any data type, matched with the right data store – NoSQL, relational, HDFS – to meet the needs of any

use case. Use cases include being able to proactively monitor, predict, and optimize IoT, cybersecurity, and digital network environments by quickly detecting and understanding abnormalities. A clear advantage is linear performance scalability with processing flexibility for developers. All of these capabilities help to extract real-time insights from big data.

Data Engineering

Cloudera Enterprise is massively scalable for any size data and its processing. The platform delivers large-scale processing to run data pipelines and power ETL and machine learning workloads. The Altus cloud-native platform-as-a-service (PaaS) for data engineering offers fast and easy execution on-demand, while on-premises and cloud infrastructure resources can be combined too. The solution brings developer productivity with the leading procedural, batch SQL, and stream processing engines, including Apache Spark and Apache Hive, Hive-on-Spark, or Apache MapReduce. It creates a foundation to train ML models with continuous ingest and processing of limitless raw data. This combination of capabilities lets you streamline and simplify your big data processing, even for continuous high-volume streams of data.

Integrated Search

Cloudera Enterprise helps you access and analyze unstructured text data via your favorite BI application, and enhance your data warehouse by enabling best-match text analytics. Most organizations only use <1% of their unstructured data (text, voice, image, genomes) for decision making, but integrated search analytics will open up the untapped value and insights of your text data. Most users and decision makers in an organization are non technical. To make your organization data driven you need to serve these users with simple ad-hoc query and analytics engines, that allows for natural language queries. Search will allow these new users to be productive on your insight hub. Other applications include automatically categorizing incoming logs events and detecting anomalies or threats in real time. Many businesses choose to o load costly and siloed log search solutions and replace with Cloudera Enterprise's integrated functions.

Meeting Enterprise Operational Needs with a Modern Platform

SDX is a set of shared open platform services built for multi-functional, multi-tenant, and/or multi-disciplinary analytics that have been optimized for the cloud. This means that Cloudera Enterprise offers a unified security model that helps protect sensitive data with a consistent set of controls, and that it o ers a consistent governance model that enables self-service secure access to all of your relevant data. Not just one type of data, really to all of it, increasing your ability to be compliant, particularly in a regulatory environment.

Data Catalog

Common pain points across analytics functions include users not being able to find relevant data, or trust what they've found, or be able to access it without IT help. Sometimes they don't know the lineage or history, sometimes the data is missing business context. Working with data in transient environments in the cloud can be particularly challenging, as this information can be lost and will need to be re-created again. A common scenario might be, "I see 10 tables called 'Customer Accounts' - which one should I use?" SDX help you identify the table that is most popular, used by all your team members, or characterized by other important attributes.

The SDX shared data catalog helps to define and preserve the structure and the business context of all your data, regardless of where it happens to reside, spanning on-premises, cloud object stores, structured, unstructured, and semi-structured data. Business catalog services (not just a Hive metastore) span all enterprise data sets, schemas, collaborative tags, and

business classifications, and are targeted for each type of user. This is enhanced by key features such as technical metadata separation, typing, and validation, and automated policy-based definitions. Persisting this information, even for temporary cloud environments makes everyone's life easier.

Security and Governance

Another set of problems with traditional and alternative approaches is around security and governance. Pain points here include incomplete or inconsistent controls, which lead to significant financial and reputation risk. Trying to solve these through administrator effort alone is burdensome, and likely won't meet industry and government regulations like GDPR, PCI DSS, or HIPAA.

Security should be an enabler, good security makes it easier to share, and avoid producing copies of data that are either stripped down or unsafe. You might be thinking, "when I add a new security policy, I need it to immediately take e ect for all workloads because my users use a mix of Apache Impala, Hive, and Spark." Great, SDX provides that, too.

Alternately, if security is too hard to configure, my users just won't use it - they'll find a way to turn it off. SDX provides automatic configuration for encryption at rest and wire encryption. You can keep key management safe in your own facility. Audit logs are complete, immutable, and preserved, unlike a cloud provider that does them at 5 minute intervals and discards after two weeks, or a Hadoop distribution that allows them to be disabled.

Cloudera Enterprise has a full complement of security features for compliance including encryption at rest and in motion, authorization by role, audit logs as noted, visibility to classify data by sensitivity, and full record updates or erasure upon request.

Lifecycle Management

Lifecycle management, often alongside partners, increases user productivity and boosts job predictability, and includes functions like flexible data ingest and replication. This is supported by a control plane handling multi-environment and multi-tenant cluster provisioning, deployment, management, and troubleshooting, described below.

As data silos cause so many problems already detailed, SDX is really a core piece of how Cloudera separates from the legacy and unintegrated competition.

A Hybrid Open Source Software Model

Cloudera's own Doug Cutting was one of the inventors of core Hadoop, but Cloudera Enterprise goes way beyond that heritage. Cloudera has a hybrid open source software model that operates in five ways

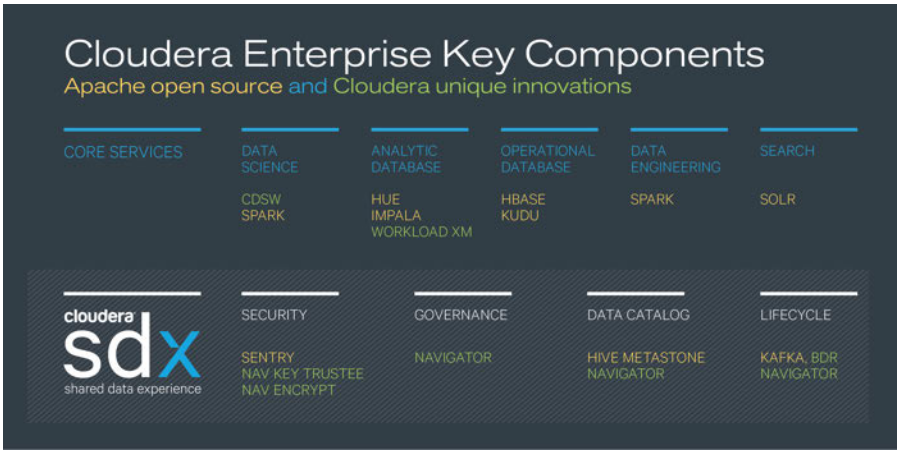
- Invent new technologies and contribute to the community
- Innovate to meet enterprise operational requirements
- Invest in testing and maturation of components
- Integrate projects into a unified, rock solid platform
- Iterate and backport for optimal blend of safety and cutting edge

For example, Cloudera has created new offerings like Impala, Kudu, and Sentry to help enterprises achieve their goals. Cloudera carefully evaluates project compatibility, not always just pushing the latest code and leaving it to the customers to figure it out, as some competitors do. When a feature is attractive but maybe not a whole component version, we sometimes backport it to more stable versions of the project. One more recent example is the

HMS separation project, where we worked with the community to pull the catalog out of the compute layer, which now underlies some of our SDX functions.

Because Cloudera operates as an open platform in an open ecosystem, we have partners spanning resellers, ISVs and solution specialists, SIs, and of course cloud and infrastructure partners, too. Some offer appliances, some complementary technology, some services, and some cloud or operating environments. Together, anything is possible.

You can see how we blend the best of open source software components and unique Cloudera innovations in the diagram below.



Key Platform Design Goals

Across all environments Cloudera's goal is to embrace a set of directional concepts and build in capabilities to the platform to support them. These include:

- Offering multi-function analytics on shared data
- Sharing more data in fewer environments
- Improving security and governance
- Reducing infrastructure costs
- Improving tenant isolation
- Improving elasticity and agility
- Improving safe self-service
- Reducing on-boarding time

These not only improve the IT infrastructure and operations, but help across all analytics teams and LOBs. Cloud services o er this promise, but Cloudera is best positioned to deliver. Private cloud will also be an increasing focus area for meeting these goals.

There is an architectural change we are driving from monolithic static to agile dynamic approaches. A “lift and shift” approach to migrating workloads to cloud doesn’t take full advantage of the possibilities, because it keeps workload, data context, policies, and compute and storage resources in one block. Simply deploying your on-premises Cloudera Enterprise cluster this same way is relatively expensive and inelastic. A better approach is separating out the SDX layer from the compute cluster, in turn enabling transient workloads WITHOUT losing context and control. Today SDX can be self-managed according to a reference architecture, or

facilitated by Altus SDX.

SDX use cases then become very interesting, and operate more dynamically. Here is an example of how multi-cluster transient workloads can run together in ways a traditional monolithic deployment wouldn't readily handle. You could combine data engineering and a data warehouse in the cloud. This enables you to run ETL, and assign permissions and classifications once, even as transient resources are used and discarded as needed. Data, along with all data context, is immediately available in the data warehouse. This creates significant agility, efficient resource utilization, and does NOT lose any important information about the data, metadata, or operations performed.

Stay tuned for more leadership as Cloudera executes creation of a universal data platform aligned upon these goals.

Deployment Model Choice and Flexibility

You may be now (re-)evaluating where to run your analytics environments. There are a lot of options, including on-premises "bare metal", private cloud, public cloud infrastructure (IaaS), and public cloud platform-as-a-service. This shouldn't be an "all or nothing" decision, and it shouldn't be a hasty one.

There are good reasons for each choice, and many businesses will end up running a mix of these options. Here are some considerations for each model:

- **Bare metal** offers cost e iciency, minimizes dependence on unproven technology, standardizes on the best ROI, and can give you complete control.
- **Private cloud** brings more elasticity and convenience, minimizes resource contention, standardizes on a single environment, and keeps things on-premises.
- **Public cloud IaaS** gives elasticity and choice of resources, minimizes dependence on you own hardware management, standardizes on one cloud, and makes it easy to scale storage and compute.
- **Public cloud PaaS** has the most agility, minimizes dependence on IT at all, focuses on ease of use for everyone, at least for a single function or analytics type.

Note again, this isn't always a migration where the goal is to end up in PaaS for simplicity, many will want a hybrid model. This does raise the question of how much of the technology stack you will have to manage at each layer. There is more control on in the top choices of the list, but more responsibility too. Options lower may be simpler, but with less ability to customize. Cloudera's goal is to o er the best platform across all four options, and we're making great progress. Not least, it's worth considering that some major cloud service providers may never have an on premises footprint, so if that's important to you, you should select a platform that can run the same way everywhere!

Cloudera Altus PaaS Goes Above and Beyond

A new focus for Cloudera to take advantage of public cloud infrastructure is Altus PaaS. Altus manages much of the cluster provisioning and administration for the IT team, so analysts can focus on their jobs, without worrying so much about resource management. Altus includes multiple workloads being rolled out over time, including Data Engineering, Data Warehouse, Data Science, and SDX. Altus functions are designed to be simple, selfservice, elastic, and role-oriented.

Altus spans multiple cloud environments including AWS and Microsoft Azure. Otherwise, it has the same capabilities and tools as Cloudera Enterprise in other deployment models. This gives you a common approach in any environment, not different platforms to chase for each.

Immediate Altus use cases could include:

- Elastic workloads with transient compute, reducing costs to only what is actively being used, as opposed to persistent "lift and shift" approaches.
- Data mart expansion where you might save on the labor and expense required to accommodate new teams and environments.
- Peak shaving where on-premises hardware is supplemented by cloud resources as needed.

Just as we have shown new agility with SDX, Altus offers new possibilities for realizing the potential of cloud elasticity and combined workloads.

You should bet on cloud providers for infrastructure, Cloudera for data, analytics, and controls, and leave the rest to the ecosystem partners. Or said another way, AWS and Azure are our partners, they like any company that brings large enterprise workloads into their clouds. At the same time, there is increasing recognition that Cloudera has more experience and capabilities to meet the needs of enterprises. And Cloudera partners are (usually) certified for cloud too, so you can use products together with confidence to enhance solutions.

Summing up Cloudera Enterprise

As shown, Cloudera Enterprise is a unique take on data management and analytics. The modern platform for machine learning and analytics, optimized for cloud, heralds a new era for businesses to transform on new insights and innovations.

What you should do next is:

1. Define your own analytics goals and initiatives
2. Explore a demo, proof-of-concept, or pilot project
3. Dive deep on the technology that makes it possible

We're here to help.

Cloudera, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com