Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

cloudera®

Embracing the "New Normal" of Big Data with Cloudera Enterprise

Putting Apache Hadoop to Work for your Organization

Table of Contents

Introduction

Rapid growth of the digital ecosystem continues to challenge an organization's ability to balance the introduction of new technology for data management methods and leveraging existing infrastructure effectively. The concept of **big data** is changing the conversation on how organizations will manage the variety, velocity, and volumes of data (the 3 Vs). Moreover, it's forcing technologists and business leaders to **infuse agility** into data modernization initiatives in order to identify opportunities to exploit data better, improve analytic capability and remove cost while remaining focused on the strategic business objectives.

Business leaders need multiple data sources to be integrated in ways that provide meaningful and practical insights, allowing them to take immediate actions. In addition to traditional data sources and business applications such as Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) systems, this increasingly means integrating external data sources such as social media, the Internet of Things, and data provided by third parties. Big data technologies empower organizations to harness unstructured, structured, and semi-structured data, much of which was previously inaccessible due to its sheer volume and the structured bias of traditional Enterprise Data Warehouse (EDW) systems.

With advancements in technology, specifically Apache Hadoop, organizations can now process petabytes of any data in a unified manner. This removes the barriers associated with traditional business intelligence platforms, and leaves organizations better equipped to rapidly process and manage the 3 Vs of data.

However, technology alone doesn't solve the big data challenge because companies struggle with adapting and evolving to **the "new normal" driven by "big data"**. Big data technologies are coming into mainstream usage and now have the industrial capabilities that enterprise data center managers look for, but the adoption rate has not been as strong as the buzz. This is in part due to the inertia around legacy technology, architecture and thought processes, and a lack of clarity around how to connect the concept of big data to strategic business initiatives to deliver value.

Business stakeholders and CIOs do look for creative ways to extract maximum value from their current data warehouses, but because these are not ideally suited to the new normal they miss out on the opportunity to add value to the business and take cost out of operations. Capgemini and Cloudera offer big data solutions that help organizations **achieve significant additional value from their data** in return for a relatively low capital investment. These solutions bring clarity to the boardroom for big data initiatives, and simplicity to the business and IT organizations.

Industry Challenges for Analytics, and the New Normal of Business Intelligence Architectures

Highly regulated industries, such as energy, utilities, healthcare and telecommunications, are being mandated by law to increase the volume of historical data they store. Other industries, such as high tech, retail and manufacturing, have similar data access, storage, and retrieval requirements because near real-time access to all relevant data provides sustainable competitive advantage. In either case, **the ability to obtain greater value from your data assets can be the difference between success and failure**.

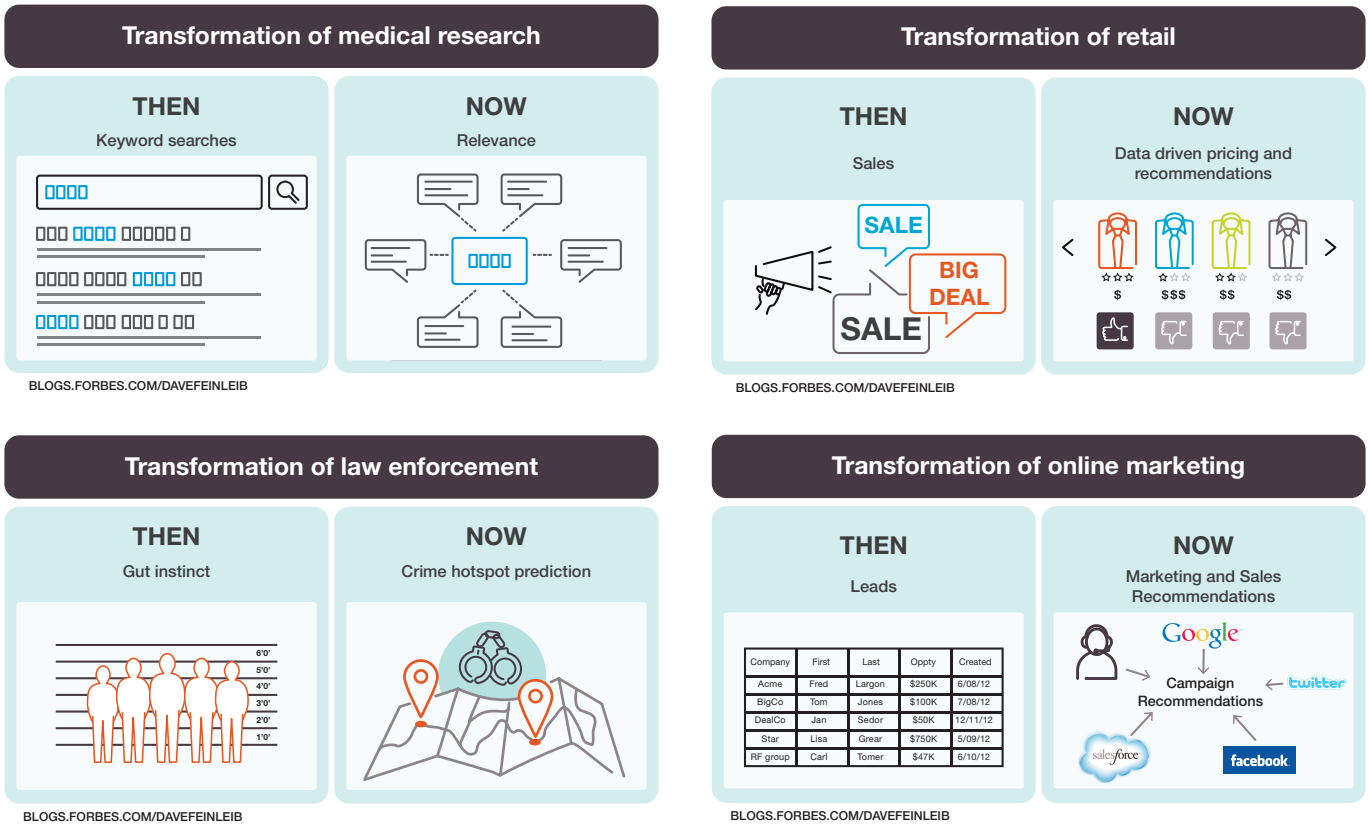
Industries are struggling with increasing data volumes, increasing complexity of data, the need for highly complex analytical models, and the pressure to manage software and hardware costs effectively. They also need to deal with the fact that around 80% of the data created today is unstructured data. These pressures, all related to big data, force organizations to evaluate how to optimize their technology landscape to gain the full value of their most strategic asset – data.

In the new normal, organizations are being pushed to leverage all the available data to make specific and personal ("segment of one") interactions with their customer base. Forbes recently published some examples of how better data access and analytics can improve decision-making in multiple industries (c.f. Figure 1).

While organizations are being pushed to evolve their business models, there is often tremendous inertia preventing or delaying progress to the new paradigm. **Such organizations run the risk that more nimble and creative players will bypass them and capture the market**.

This article describes some of the reasons why organizations are still reluctant to jump into big data and how Capgemini and Cloudera Enterprise solutions can help overcome their hesitation and get started now.

Figure 1: Transformation in selected industries



Challenge 1: Inflexible structure and lack of business agility

By nature, data warehouses are structured systems. They are designed to present data in predetermined ways as defined by a data model and the requirements set out by the business teams managing the data warehouse.

This concept, called "pre-modeling", is designed to serve only predetermined needs and use cases. Pre-modeling makes it very difficult for business users to work on hypotheses that have not previously been considered, unless they embark on the traditional IT software development life cycle. Further, the structured nature of the data warehouse prevents businesses from being able to add context and value from information available in unstructured form such as documents, contracts, call notes, social data, web logs, etc.

Although data warehouses provide organizations with substantial value on a daily basis, they severely constrict exploratory use cases, innovation, and "fail fast" initiatives. The inability to use all your data for analytics can mean missed innovation opportunities or loss of competitive advantage.

Challenge 2: Industrialization, prioritization and shadow IT

Legacy data warehouse systems are often designed like the ERP systems alongside which they were meant to operate. The data warehouse is now often designated a "production" system, with SLAs for predetermined functions. This doesn't leave much room for experimentation and hypothesis building.

To get around this limitation, you often see shadow IT groups trying to work around the constraints that an ERP-centric IT department has placed on them. In some cases, business users claim their data back, and it's not uncommon to find hundreds of MS Excel extracts from a data warehouse on business users' laptops. This undesirable situation poses great challenges in terms of data governance, quality and security.

Challenge 3: Cost per terabyte: breaking though the new glass ceiling

With performance and speed of analytics getting faster, the new limit organizations are currently facing is actually about storage and computing costs. Some organizations pay up to $35,000 per terabyte on an annual basis (see World Quality Report 2013-14 | Capgemini Worldwide).

With new business requirements coming in daily and data sets growing faster than ever, companies are forced to restrict their IT spending. IT is already saying "I just can't spend any more money this month/quarter/year". Now, imagine today's data requirement multiplied 10 or 100 times, and throw in the curve ball of unstructured data, which as we have seen represents 80% of all data.

To remain within their IT budget, companies often need to come up with capability-limiting workarounds such as:

- Archive older/"colder" data (e.g. last year's CDRs for telcos) that is too expensive to store given that the value they will bring is unknown.
- Postpone the delivery of new business requirements (new KPIs, new analytics, etc.) as the current architecture will not support the corresponding workload.
- Restrict exploratory use cases and related resources, as new extensions to the current platform would become mandatory.

Organizations are often forced not to retain all their data in the warehouse, and instead to archive it to more affordable offline systems, such as a storage grid or tape backup. A typical strategy is to define a time window for data retention, beyond which data is archived. Others choose to aggregate the data or sample it.

Either way, business users and analysts can't draw insights from data that is not in the warehouse. Often, only acts of God (data center destruction) or government requirements can cause data to be brought back from tape archive into the warehouse for analysis.

Challenge 4: The pitfalls of data modeling "on write"

Within a data warehouse, the data modeling of the detailed data layer is a critical point. It helps the organization align itself around key concepts and entities like customers, contracts, assets, catalogues, and usage data to ensure that the entire company and business user population speak the same language. It helps provide a robust "single version of the truth" on which business departments can build their KPIs.

But this advantage comes with constraints, because the data is transformed and integrated into this data model as soon as it is brought into the data warehouse ("on write"). Companies can choose to keep the "raw" version of the data for a short period, but once again the cost challenge kicks in. Until now, there has been no cost-effective way for organizations to keep as much raw data as they would like to.

Therefore, when new attributes need to be added because of a change of a data source interface (e.g. an ERP upgrade) or if new business processes are implemented (e.g. new rate plans, new product catalogue), organizations have to go through significant changes in the data model and data migration processes before being able to use this new data.

Using Hadoop as a data reservoir helps companies take advantage of the concept of data modeling "on read": that is, defining the data model at the time of deriving intelligence from data. Instead of tweaking and transforming the data before storing it, companies can define a core data model

when the raw version of the data is read. This enables them to be much more agile when they need to modify the data model, as they no longer need to migrate data.

This doesn't mean that there is no need for data modeling – organizations still need to understand data to be able to use it. However, defining the data model "on read" does mean that organizations can drastically improve time to market when they need to modify the view chosen for the data.

Challenge 5: Agility on workload diversity

Current data architectures do not allow much agility in terms of serving the ideation needs of clients while also accomplishing compute-intensive transformations for production purposes. Some data warehouses still struggle with the problem of mixed workload prioritization, given that their ability to handle bigger and more complex datasets is limited by their compute and disk capacity.

Extending these capabilities with an Apache Hadoop platform, which is an order of magnitude cheaper and implemented on industry-standard hardware, allows businesses to bring massive parallel compute and disk capacity to bear. Organizations can then run many different kinds of workloads at a fraction of the cost, and allocate resources to deliver private/department-specific studies, exploratory analytics, and benchmarks of new KPIs or analytical models. As these exploratory workloads will probably increase exponentially, the right governance needs to be set up between business and IT teams in order to capitalize on these initiatives, ensure they fail fast when appropriate, and industrialize the successful ones.

Challenge 6: High availability and disaster recovery

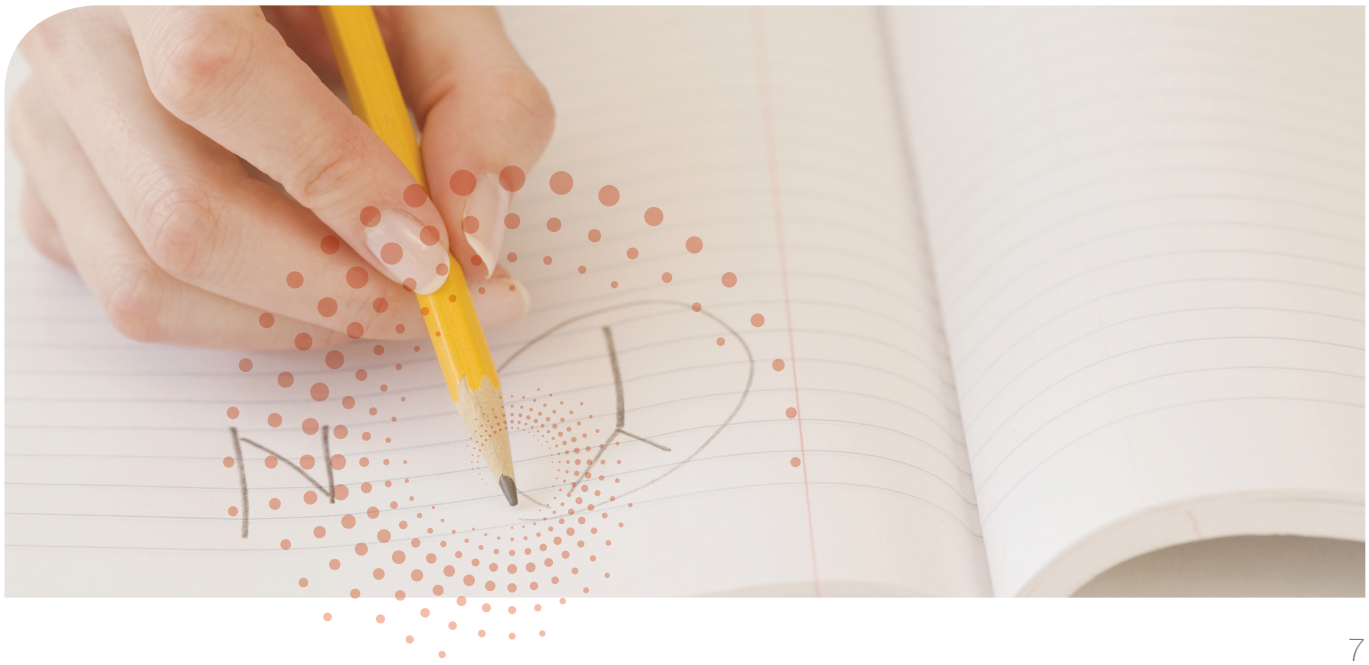
Until just a few years ago, data warehouses were the de facto solution for all things analytic in an organization, but were typically implemented in a "back office" mode. However, since the "information explosion", Information Management has become critical not merely to operations, but to the viability of the business. For many of our clients, information management systems are now on the critical path for business operations, and have uptime requirements comparable with those of ERP and CRM systems.

While business criticality has changed, IT budgets have not seen a corresponding uplift. Companies still struggle to afford duplicate high-cost EDW systems that meet strict fail-over and disaster recovery (DR) requirements.

Challenge 7: Capacity constraints for reporting and analytics

The evolution and refinement of extract, transform, load (ETL) and extract, load, transform (ELT) has permitted a drastic optimization of the performance of daily batch windows. However, the transformation portion often takes up to 40-60% of the available compute capacity of expensive data warehouses, while staging takes up to 30-50% of their storage capacity. As data sets grow, ETL/ELT processes are challenged on SLAs.

This situation leaves ever more limited resources available for structured reporting and analysis, let alone for ideation and experimentation on data. The resulting constraints often lead business units to seek solutions outside of the IT BI group, or give up on their BI quest because of lack of capacity, or for cost reasons.



Capgemini's Recommended Reference Architecture

To help organizations evolve to the new normal, Capgemini recommends that they augment their existing data warehouse platforms with Cloudera, the big data platform built on Apache Hadoop. Cloudera helps solve the challenges described above by:

- Providing cost-effective data storage and computing power, allowing such architectures to become practically limitless in terms of size of data sets.
- Providing for the capture and use of broader and richer data sets including machine-to-machine, social, mobile, network, etc.
- Allowing bigger and more complex questions to be answered, using new advanced analytics techniques like R.
- Making the use of data throughout the organization more pervasive and agile by providing a SQL query engine together with Search capabilities.
- Providing an enhanced user/analyst experience with visualization tools.

The following diagram illustrates Capgemini's high-level Information Management Reference Architecture and Hadoop's role in the Enterprise Data Platform.

In this high-level architecture, the Hadoop platform serves multiple purposes. It provides:

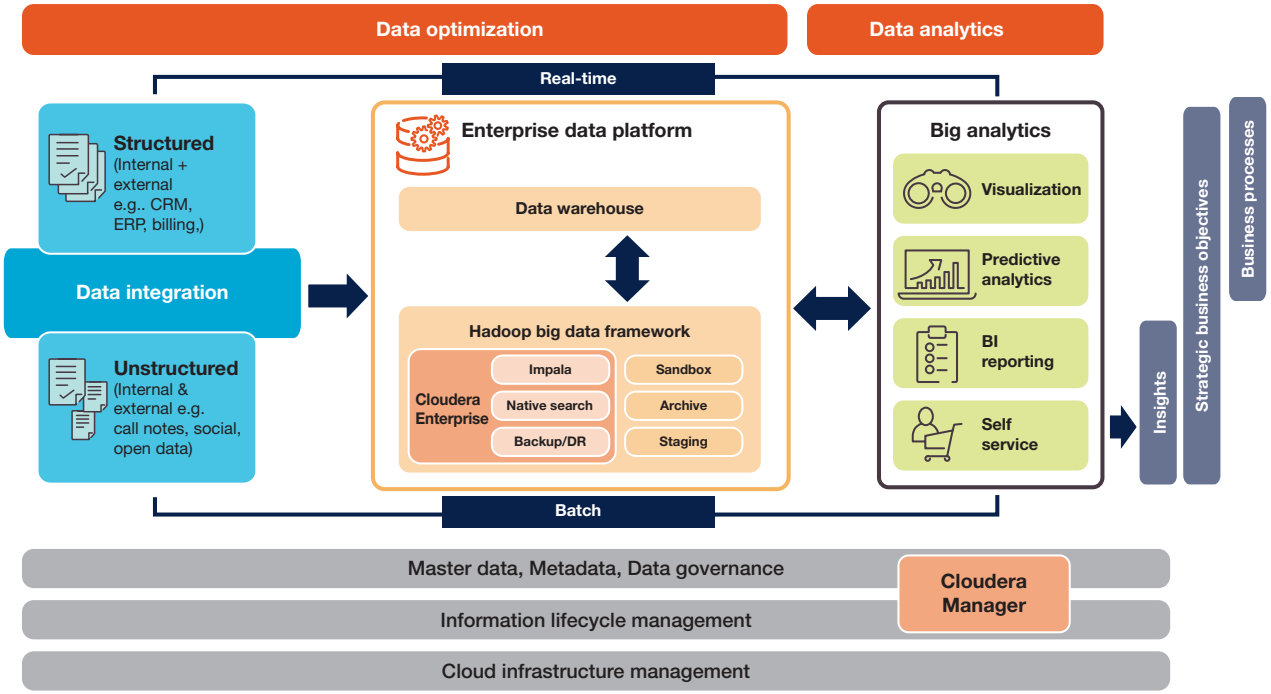
1. The staging repository for all data – raw, interim and processed – with as much history as needed, and the ability also to store all unstructured/semi-structured data, and structure it at will.
2. Low-cost, high-efficiency storage, processing, and multi-faceted data modeling.
3. An Ideation Analytics Sandbox for hypothesis building and verification, specific business studies, and benchmarks.

To make sure companies get the maximum possible value from this new type of architecture and from their data, an end-to-end enterprise information governance process that leaves room for business agility is a key factor of success.

Cloudera as the Enterprise Data Hub

When it comes to data management, Cloudera's data-centric, schema-on-read architecture (defining the data model at the time of data retrieval) allows organizations to store all their data – regardless of type – in its original format, without the need for extensive data transformation.

Figure 2: Analytical architecture complemented by Cloudera



Some data flows will have Cloudera Hadoop process raw data and then populate the data warehouse with high-value, cleansed, and conformed data. For some other data sets, data will remain in Cloudera Hadoop and analytics will be done directly, without moving or replicating data into the warehouse.

Fundamentally, Cloudera aacts as the data hub, or reservoir, that can scalably, flexibly, and economically store, process, analyze, and serve a variety of information use cases. This allows the data warehouse to focus on the reporting and analytical function that it was purchased for.

Cloudera Hadoop as the active data archive

The primary function of the traditional data warehouse or data mart is to facilitate regular, sometimes real-time, structured reporting for pre-defined use cases and business requirements. Cold data sets that are not needed for immediate operational purposes, or are used less frequently, are often offloaded to near-line or offline storage.

By extending the architecture with Cloudera, this cold data can now be moved into the Hadoop platform so that it can be analyzed as often as needed. Cloudera Hadoop is a cost-effective and reliable alternative to tape storage, making longer-term analysis of high-volume data straightforward. This enables deeper and more accurate analysis, as you do not have to draw conclusions from insights gained from partial or aggregated data sets.

Cloudera Hadoop as the ideation platform for structured and unstructured data

As we have seen, Cloudera becomes the organization's data hub, with low-latency SQL & BI support and the ability to query across all data with tools such as Cloudera Impala, or do full text Search with Cloudera Search. The business's power to test out hypotheses then increases dramatically because it can analyze a vastly larger amount of history and can work across structured and unstructured data.

Traditional data warehouses were constrained by pre-modeling that forced data into a single pre-determined structure suited to a predetermined format. Hadoop, by nature, allows data to be stored in its raw format and used whatever way is required, and in many different ways at once. We call this approach "post-modeling" or "modeling on read": it gives you the ability to build and adapt data models at will, for a specific use case or hypothesis.

Post-modeling, together with the ability to query across structured and unstructured data, massively increases the value users can get from data.

Offloading data processing (ETL/ ELT) to Cloudera Hadoop

By shifting their data processing load to Hadoop, enterprises can simultaneously accelerate their ETL/ELT pipelines and alleviate resource contention stemming from processing on existing data warehouse systems. With Cloudera, enterprises enjoy greater performance throughout the complete ETL lifecycle. They also gain a broad range of processing capabilities, including larger source storage, flexible and scalable staging needs, and batch and short-cycle processing.

In addition, integrating semi- or multi-structured data sets within a standard relational data warehouse, using existing ETL tools or ELT processes, is a quite complex task. The choice is sometimes made to not exploit such types of data because of the complexity that would result.

Hadoop's low cost and its linear scaling approach allow the deployment of solutions where data can be ingested at low cost and made available for querying and search via a variety of data models and patterns. This approach greatly improves analysis and reporting capabilities. Offloading data processing to Cloudera Enterprise allows the organization to use the data warehouse's compute and disk capacity for the high-value analytics functions it was purchased and built to perform.

Cloudera as extended analytical capability

Because of the volume of data that can be stored on Hadoop, and the fact that statistical languages and tools such as R and SAS can use that data natively in a distributed fashion, companies can now use the true power of a massively parallel system for deep statistical analysis. It also becomes possible to perform full-fidelity analysis: that is, analysis across the entire scope and breadth of the relevant data sets, directly on detailed data.

As a result, models avoid fundamental statistical errors due to sampling. Users no longer have to obtain a small sample of data and run it on an underpowered statistical desktop or server. Instead, they run full-fidelity analytics, on complete data sets, in a fraction of the time it took to run these analyses on sampled data.

High availability and disaster recovery with Cloudera

Given the order-of-magnitude cost differential between Hadoop and traditional appliances, it becomes easy to achieve IT "hardening" to meet demanding high availability and Disaster Recovery goals. The following features – available natively with Cloudera Enterprise - allow the Information Management System to meet critical SLA requirements at a fraction of the usual cost:

1. Native high availability (data is typically replicated onto three nodes)
2. Disaster recovery process allowing periodic refreshes of DR system
3. Downtime-less upgrades for non-major releases
4. Auditability

Why Cloudera?

Cloudera provides a complete, tested, and widely deployed open source distribution of Apache Hadoop, offering batch processing, interactive SQL and interactive search, all complemented by an enterprise-grade administration software suite.

Capgemini has partnered with Cloudera to make Apache Hadoop capabilities available for mainstream adoption by its clients. Capgemini has embedded specific Hadoop capabilities and skills into its existing frameworks for Business Information Management.

Open source, real-time query for Hadoop with Cloudera Impala

- Perform interactive analytics directly on data stored in Hadoop without the bottlenecks caused by data movement and by jumping between data silos.
- Reduce data movement and remove duplicated storage by performing interactive analysis directly on full-fidelity data.
- Leverage existing BI tools and employee skill sets (e.g. SQL) to interact with data stored in Hadoop.
- Enable more users to interact with more data by providing a single repository and metadata store from source to analysis.

End-to-end administration for Hadoop with Cloudera Manager

- Deploy, configure, and operate clusters with centralized, intuitive administration for all services, hosts, and workflows.
- Maintain a central view of all activity in a cluster through heat maps, proactive health checks, and alerts.
- Diagnose and resolve issues using operational reports and dashboards, events, intuitive log viewing and search, audit trails, and integration with Cloudera Support.
- Integrate Cloudera Manager with existing enterprise monitoring tools through SNMP, SMTP, and a comprehensive API.

Real-time search with Cloudera Search

- Help analysts and non-technical users find relevant data through user-friendly search and drill-down navigation across large, disparate data stores with mixed formats and structures.

Cloudera's industry-leading Big Data platform features

- **Powerful**
  Store, process, and analyze all your data to drive competitive advantage
- **Efficient**
  Hadoop unifies compute and data to improve operational efficiency
- **Open**
  100% open source: CDH is the world's most popular open source distribution powered by Apache Hadoop
- **Simple**
  Easy to deploy and operate, with centralized administration
- **Compatible**
  Leverage your existing investments for rapid adoption and lower TCO
- **Economical**
  Rethink the economics of data management with an open source platform on industry standard hardware – up to 90% more cost-effective than traditional solutions
- **Enterprise-ready**
  Equipped with critical capabilities to support mission-critical operations including enterprise-class high availability and Disaster Recovery (DR) capabilities

- Discover the "shape of data" quickly and easily during data modeling and exploration with faceted search interfaces and free-text query APIs.
- Consolidate data silos and minimize expensive data movement by sharing source data with multiple computing frameworks on the same Hadoop cluster, including search.

Open source, fine-grained access control with Cloudera Sentry

- Ensure the right resources have the proper and relevant permissions for appropriate data or subsets of data and SQL activities in Hive and Impala.
- Simplify administration by granting sets of permissions to resources within the organization based on functional roles within a Hive or Impala database.
- Store sensitive data alongside non-sensitive data in the same data set within Hadoop without replication, and ensure usage and data complies with regulations and governance policies.
- Empower new and varied users, and facilitate the use of a range of data, within the enterprise, alleviating security concerns by building on the foundations of concurrency, authentication, and authorization provided by Hive, Impala, and Sentry.
- Build multi-user applications on top of Hive and Impala by segregating access to data sets for appropriate users and delegating permissions management to local database administrators.
- Avoid suboptimal choices for authorization– for example, self-regulated, "benevolent" advisory authorization or "all-or-nothing," coarse-grained, file-based access.
- Build on existing systems like the Hive meta-store and establish a solid, open, and extensible framework for fine-grain authorization and security beyond SQL on Hadoop.

Data audit and access management with Cloudera Navigator

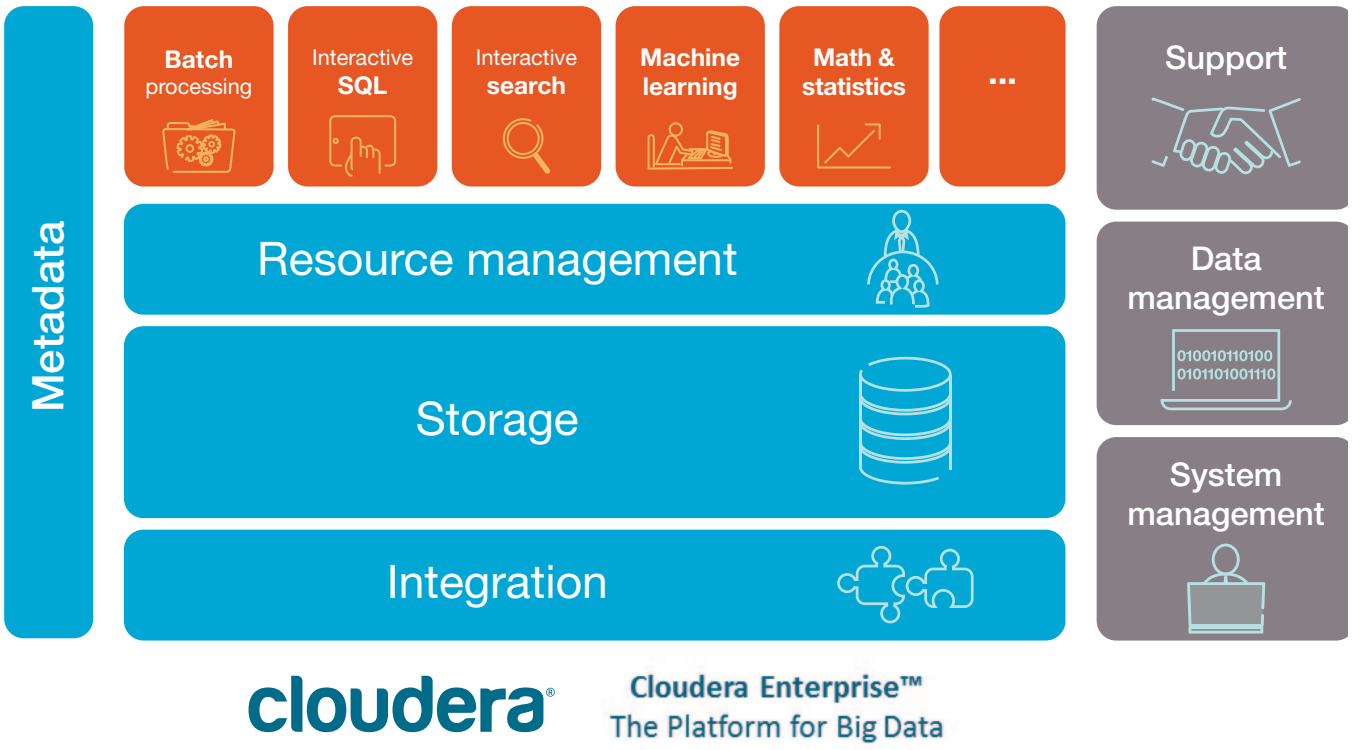
- Verify access permissions to files and directories.
- Maintain a full audit history of HDFS, Hive and HBase data access.
- Report on data access by user and type.
- Integrate with third-party security information and event management (SIEM) tools.

Enterprise Support

Cloudera offers the industry's highest quality technical support for Apache Hadoop, with a dedicated team of support engineers comprised of contributors and committers for every component of CDH, the market-leading open source Apache Hadoop distribution.

With Cloudera Support behind you, you'll experience more uptime, faster issue resolution, and better performance to support your mission critical applications.

Figure 3: Cloudera Enterprise™ Platform for Big Data



How to Make Big Data Work for Your Organization

Capgemini has extended its Business Information Management service model to leverage the new capabilities brought by these new normal architectures, and to make sure we put the right skills and expertise to work for our clients.

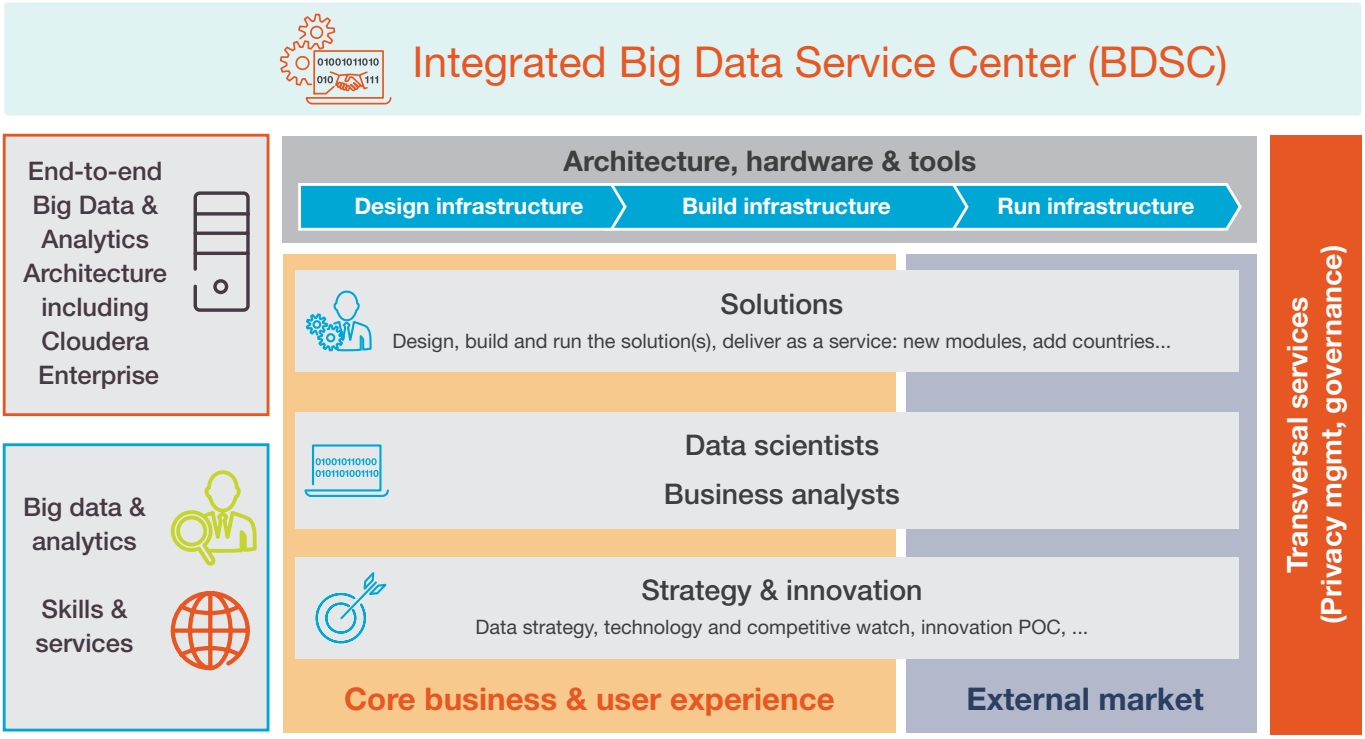
We know our clients will have new questions for us as a result. You may be thinking:

- What is the impact of these new technologies on my landscape?
- How do I guarantee more agility and more satisfaction for my business users?
- Do I risk ending up with a "data black hole", where using data is even more complex than with the traditional databases that we understand?
- How do I guarantee that I will meet my SLAs, comply with production systems regulations, and fulfill security requirements?
- How do I make sure I will have the right skills to operate and manage this new technology?
- Will I incur hidden costs because I'll need to hire high-end expertise?

These are exactly the type of questions that Capgemini and Cloudera together can help solve. By partnering with Cloudera, Capgemini aims to bring Hadoop capabilities into the mainstream. It will integrate Cloudera's best-of-breed products, support and expert services with Capgemini's own capabilities and services offerings for Business Information Management.

In collaboration with its clients, Capgemini has built a framework called the Big Data Service Center (BDSC). The objective is to help organizations deliver "Big Data Value" leveraging Capgemini's Rightshore® approach. The framework enables the definition and set up of governance, processes and agile organization for our clients' business and IT teams. It helps them to make big data work by providing an industrialized, low-cost, high-performance big data delivery and support capability.

Figure 4: Big Data Service Center Framework



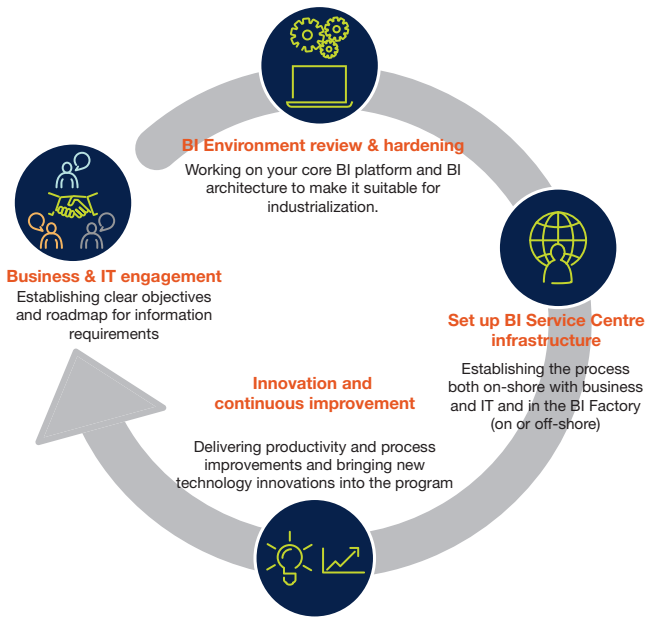
Capgemini's Big Data Service Center (BDSC) framework makes it possible to achieve business transformation through big data. This approach aims to deliver better, faster, more reliable, more business-relevant information and insight to business users by providing:

- The people, processes and governance to engage with business and IT to deliver meaningful insights where and when needed
- Low-cost, high-performance, industrialized big data development and project support, with Rightshore® leverage
- A scalable big data development factory (offshore, near-shore and on-shore).

Capgemini has extended its existing Business Information Service Center (BISC) approach to include the specific skills needed to implement the extended architectures described above.

When using these advanced architectures, it is particularly important to optimize well-known areas like data governance, data quality, security, and functional and technical administration. Capgemini believes that only a comprehensive, end-to-end service framework, encompassing all work streams necessary to deliver value to both IT & Business teams, can fully deliver on the big data promises.

Figure 5: End-to-End Service Framework



The key work streams and capabilities enabled by Capgemini's Big Data Service Center are:

- **Demand management:** A clearly defined process for enabling development of 100s of work packages (if required) across multi-stream programs.
- **Open work package costing:** The Capgemini GREAT BI tool provides a clear and auditable basis for work package costing. Organizations can easily see what they are getting and how productivity improvements will deliver cost savings over time.
- **Flexible resource management:** Using its BIM Center of Excellence resource pool, Capgemini is able to scale projects up rapidly to meet business needs, and similarly scale down again when demand reduces.
- **Program stream and multi-vendor engagement:** Capgemini's program management methodology is designed to deal with multiple work streams and multiple vendors' development teams.
- **Expertise bureau:** Having ramped up new capabilities specifically for Cloudera Hadoop, Capgemini is able to bring the right level of expertise to the delivery team, ensuring top-quality implementations. Capgemini's close partnership with Cloudera means that you also have immediate access to Cloudera's expertise.
- **Central design authority:** Providing consistency in design across multiple streams of work.
- **Service introduction:** Ensuring smooth transition with no business disruption.
- **Data analyst SWAT teams:** An objective of Capgemini's BDSC is to help find new "golden nuggets" in clients' data sets, or in third-party data sets that they may acquire and cross-analyze. We'll set up and deliver specific use case discovery task forces: SWAT teams of data scientists who work within the project team, and with clients' business users, to discover the specific competitive advantage that's most relevant for them.
- **Strategy & innovation:** Because Capgemini sees big data as a powerful transformation lever, we have integrated a specific strategy & innovation work stream into our BDSC. Our objective is to help clients define, drive and adjust their big data strategy, taking into account new insights gained from the extended enterprise data platform, and shaping innovative initiatives like new business models around, for example, data monetization.

To get you started, Capgemini uses its big data capabilities to help you decide on the first steps to take, and the way forward for your teams. Our flexible, innovative options for exploring the big data technology landscape include:

- Capgemini's **Elastic Analytics** offering: This leverages Amazon Web Services (AWS) to enable use case functionality rapidly and to provision technology in the cloud using a variety of solutions including Cloudera. Clients can create an "art of the possible" environment that demonstrates capability, or can use the platform in a production capacity. They need incur no infrastructure capital expenditures to establish and run the environment. The offer's elasticity ensures that it expands as the client's information and capability needs grow, and terminates when the work is complete.
- Capgemini's BIM **CUBE** (**CU**stomer **B**IM **E**xperience) innovation lab implements technology that creates use cases, together with Rapid Design Visualization techniques, to help clients explore what technology could deliver "in real life".

To help define the way forward and plot the relevant business & IT transformation roadmaps, Capgemini has developed a methodology called Data Optimization with Cloudera. Through a Strategic Value Assessment (SVA) with a client's teams, Capgemini determines how to implement a viable big data strategy that will to help the organization manage all its data as a strategic asset. The SVA identifies potential costs reduction opportunities coupled with improved value-based analytics.

Summary

You may already be testing Hadoop in-house, or perhaps you have yet to get started. Maybe you've already identified the "golden use cases" for your big data implementations – or are you still exploring what's possible?

Whatever stage you're at, Capgemini and Cloudera together have the ability to help you define and execute your big data strategy.

Today, everybody's talking about big data. By working with Capgemini and Cloudera, you can derive true business value from it.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

cloudera®

About Cloudera

Founded in 2008, Cloudera pioneered the business case for Hadoop with CDH, the world's most comprehensive, thoroughly tested and widely deployed 100% open source distribution of Apache Hadoop in both commercial and non-commercial environments. Now, the company is redefining data management with its Platform for Big Data, Cloudera Enterprise, empowering enterprises to Ask Bigger Questions™ and gain rich, actionable insights from all their data, to quickly and easily derive real business value that translates into competitive advantage. As the top contributor to the Apache open source community and leading educator of data professionals with the broadest array of Hadoop training and certification programs, Cloudera also offers comprehensive consulting services. Over 700 partners across hardware, software and services have teamed with Cloudera to help meet organizations' Big Data goals. With tens of thousands of nodes under management and hundreds of customers across diverse markets, Cloudera is the category leader that has set the standard for Hadoop in the enterprise.

www.cloudera.com

contact us at
partner@cloudera.com

About Capgemini

With more than 125,000 people in 44 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2012 global revenues of EUR 10.3 billion.

Together with its clients, Capgemini creates and delivers business and technology solutions that fit their needs and drive the results they want. A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

For further information visit
www.capgemini.com/bim

or contact us at
bim@capgemini.com

MCOS_GI_MK_20131021