accenture

Strategy | Consulting | Digital | Technology | Operations

Building a modern data supply chain to accelerate data

A reference architecture designed for data acceleration leveraging SAP HANA and Cloudera's Big Data Platform

High performance. Delivered.

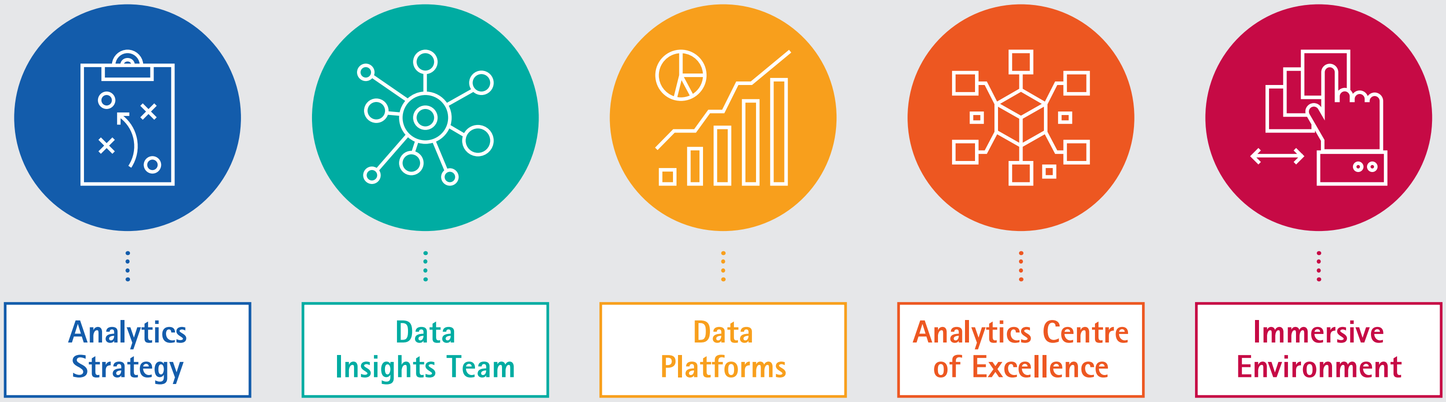
Overview

Data technologies are evolving rapidly, and organizations have adopted most of these in piecemeal fashion. As a result, enterprise data — whether related to customer interactions, business performance, computer notifications, or external events in the business environment — is vastly underutilized. Moreover, companies' data ecosystems have become complex and littered with data silos. This makes data more difficult to access, which in turn limits the value that organizations can get out of it. Indeed, according to a Gartner, Inc. report, 85 percent of Fortune 500 organizations will be unable to exploit Big Data for competitive advantage through 2015.[1] Furthermore, a recent Accenture study found that half of all companies have concerns about the accuracy of their data, and the majority of executives are unclear about the business outcomes they are getting from their data analytics programs.[2]

To help unlock the value hidden in their data, companies should start treating data like a supply chain, enabling it to flow easier and usefully through the entire organization—and eventually throughout each company's ecosystem of partners, including suppliers and customers. The time is right for this approach. For one thing, new external data sources are becoming available, providing fresh opportunities for data insights. In addition, the tools and technology required to build a better data platform are available and in use. These provide a foundation on which companies can construct an integrated, end-to-end data supply chain.

The journey to improved decision-making and analytics ROI can be complex. To help support our clients in making the end to end journey, Accenture has established our 5 pillar framework (shown below in Figure 1), the Accenture Connected Analytics Experience (ACAX). These five pillars represent support systems to help make data and analytics more accessible and engaging at all levels of the organization.[3]

FIGURE 1: Insight-Powered Enterprise



The focus of this paper is on the third pillar, Data Platforms. It is our position that tomorrow's most successful organizations will utilize multiple technologies to create their modern data supply chains and accelerate data and thus insights and outcomes.

A modern data supply chain begins when data is created, imported, or combined with other data. The data moves through the links in the chain, incrementally acquiring value. The supply chain ends with actionable, valuable business insights — such as ideas for new product, service, process innovations, marketing campaigns, or globalization strategies. Configured effectively, a data supply chain helps enable organizations to discover their data, leverage more data sources, and move data faster. These capabilities, in turn, position an organization to extract more value from its data through advanced computing techniques such as machine learning.

A modern data supply chain helps organizations surmount three data-related challenges:

Movement:

How to move data faster from its source to places in the organization where it is needed

Processing:

How to process data to gain actionable insights faster

Interactivity:

How to foster faster responses to queries submitted by users or applications

In this point of view, we will closely examine those three data related challenges and assess the landscape of architectural components available to address them. We will then explore options for combining these components to create data platform solutions.

Three challenges that the modern data supply chain should address

Movement

Traditionally, bringing data into an organization was a slow but fairly straightforward process: Data was collected in a staging area and then transformed in to the appropriate format. The data was then loaded to reside in one location, such as a mainframe or enterprise data warehouse. From there it was directly transferred in a point-to-point fashion to a data mart for users and applications to access. However, with the mammoth increase in data volumes and variety, such a traditional process no longer works effectively.

For example, the Internet of Things (IoT) is playing a major role in driving new developments in data movement. In its simplest sense, the IoT comprises connected devices — ranging from refrigerators, smart meters, and video cameras to mobile phones and children's toys — that could be located anywhere in the world. According to Gartner, Inc., there will be as many as 26 billion devices on the IoT by 2020. Every connected device generates data, each with its own format and idiosyncrasies.

Whether a business is deploying thousands of individual systems or simply trying to keep up with its own growth, having a modern data infrastructure in place that can collect relevant data can lead to differentiation by enabling data insights. But to extract valuable insights from data in this new world, organizations need to harness it from multiple sources without losing any of it, and deliver it for processing and storage. Some data exists as log files on external systems that have to be transported to an organization's data infrastructure for future use. Other sources provide streaming data, which is piped into the system in real time; that is, as the data is generated. Examples include power consumption information from smart electrical meters that is always updating.

Whatever the source and format, collecting data from its origin to where it is needed in the organization can seem like drinking from a fire hose while trying not to lose a single drop. The modern data supply chain helps organizations manage this feat by enabling multiple ways of bringing data into an organization's data infrastructure and ensuring that it can be referenced quickly.

Processing

Organizations have long been processing data in an effort to extract actionable insights from it. However, the volume and variety of data requiring processing have ballooned. To accommodate growth on those two fronts and generate faster but also accurate results, enterprises have to step up their processing capabilities. In particular, they must carry out three activities more speedily than ever: performing calculations on the data, creating and executing simulation models, and comparing statistics to derive new insights from the data.

The rise of real-time analytical technologies has presented new opportunities on this front. A good analytical technology is able to pre-process incoming data. For example, by monitoring a customer's location, an organization can deliver a promotion or discount to a customer's mobile device when he or she is near a likely place of purchase. But a better technology combines streaming data with historical (modeled) data to enable more intelligent decision-making. For instance, by correlating a customer's location with his or her previous purchase history, the company can deliver a promotion that is tailored to that same customer, increasing the likelihood of conversion.

To reap the full benefits of faster data processing, companies must make better use of computer clusters — organized sets of hundreds or thousands of computers working together to sift through large amounts of data. With the cost of random-access memory (RAM) at an all-time low, new solutions for extracting data from storage more quickly have bombarded the market, each with its own promise of speed, durability, and accuracy. Data acceleration supports faster processing by leveraging advances in hardware and software for computer clusters, enabling them to operate more efficiently than ever.

Interactivity

Interactivity is about usability of the data infrastructure. Fundamentally, users or applications submit queries to the infrastructure and expect to receive responses to the queries within an acceptable amount of time. Traditional solutions have made it easy for people to submit SQL-based queries to get the results they need to arrive at actionable insights. However, the rise of big data has led to new programming languages that discourage existing users from adopting the systems. Additionally, owing to the sheer volume of data, users may have to wait many minutes or even hours for results on a query. The longer users have to wait, the longer it takes them to gain insights required to make critical business decision. That is the case whether clients are internal (for example, a marketing director who wants to know which of the company's customers are most loyal and profitable) or external (for example, a business process outsourcing (BPO) client company that needs to know how performance of an outsourced process has changed over the term of the BPO engagement).

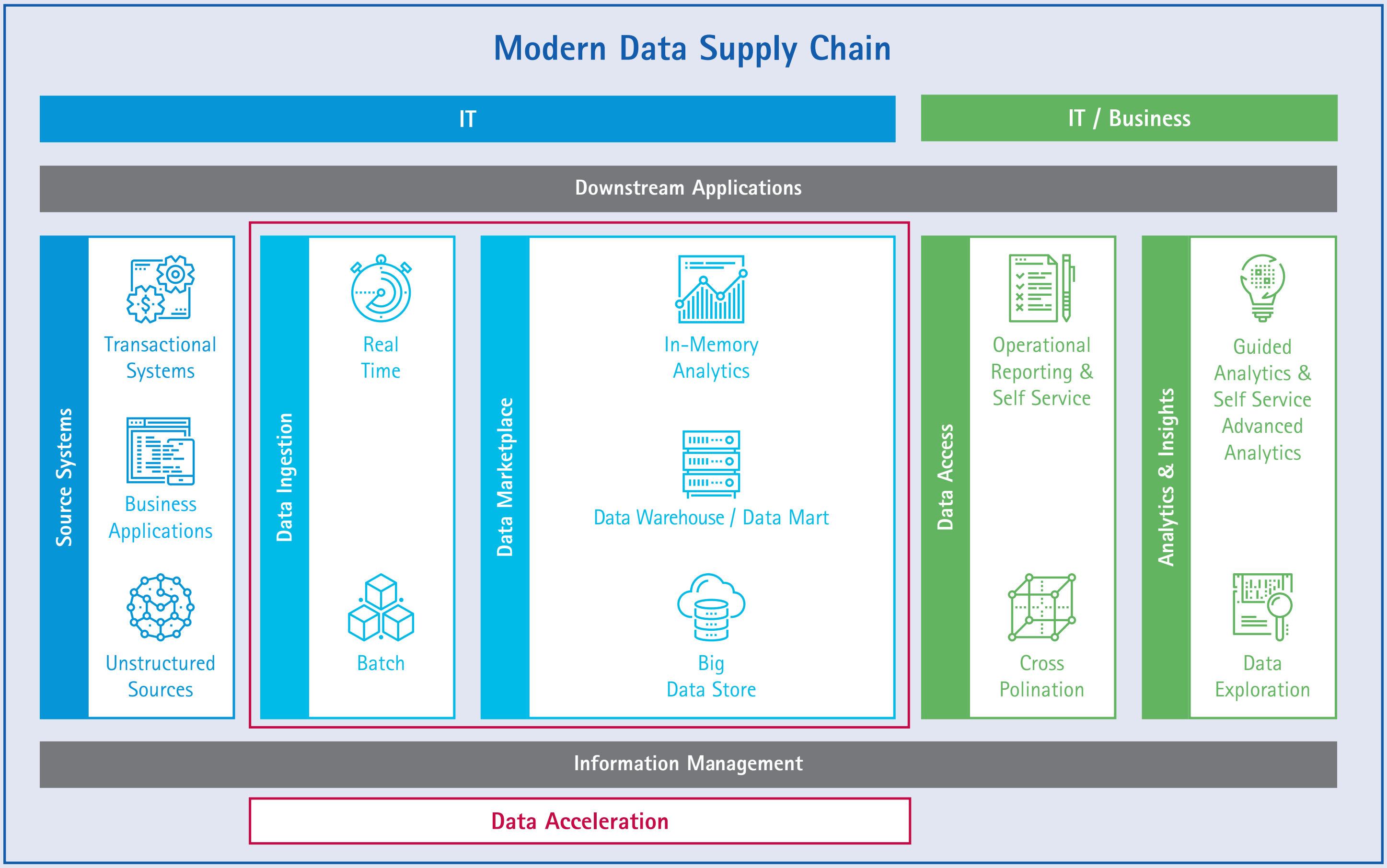
Clients providing critical services to their own customers — such as retail transaction processing — might require response times in the sub-second (millisecond) range. With less critical business tasks, acceptable response times may be longer. Modern data supply chain supports faster interactivity by enabling users and applications to connect to the data infrastructure in universally acceptable ways and by ensuring that query results are delivered as quickly as required.

Understanding the architecture landscape

Organizations can choose from many different data technology components to build their modern data supply chain and accelerate their data. These include big data platforms, complex event processing, ingestion, in-memory databases, appliances, etc. Each component can address data movement, processing, and/or interactivity, and each has distinctive technology features.

These architecture components cannot function in isolation to support data acceleration, however. Instead, they must "play well" with each other, capitalizing on one another's advantages. In the sections that follow, we take a closer look at these components.

FIGURE 2: Reference architecture for a modern data supply chain



Downstream Applications

Applications that are consuming information from the storage.

Data Ingestion

An abstract layer that allows various data elements – type and latency to flow into the Data Marketplace.

Data Marketplace

Includes the components that allow to store and transform raw data into information.

Data Access

System agnostic semantic layer that allows access to the data for the end consumption – interactive, blending, automation, low latency.

Analytics & Insights

Single point of entry that support information presentation, analysis, and advanced analytics methods to users.

Data Ingestion

Ingestion is all about collecting, capturing, and moving data from its sources to underlying repositories where users can process it. Traditional ingestion was done in an extract-transform-load (ETL) method aimed at ensuring organized and complete data. Modern data infrastructure is less concerned about the structure of the data as it enters the system and more about making sure the data is collected. Modern techniques act on streaming data, such as continuous clicks on a website, and involves queues (processing of the data in the appropriate order).

As noted earlier, organizations need a mechanism for capturing data from multiple external sources (each of which might deliver data in different formats and might have different requirements) and quickly transporting the data to a place where users can access it for processing. The data can be static and reside in a repository external to the organization's data infrastructure — or it may be generated in real time by the external source. Ingestion solutions offer mechanisms for accessing and using data in both scenarios. In this "pub-sub" system, the producer of the data publishes it from the source to a buffer or channel (data holding area). The subscriber (user or consumer) of the data picks it up from there. A queuing mechanism allows data to be buffered while the system waits for producers and consumers to take their respective actions. The speed of data producers' and consumers' actions determines the size of the buffer and the queue. Robust ingestion supports data acceleration by enabling large amounts of data to be collected and stored quickly.

Data Marketplace

A Data Marketplace is an integrated platform with big data store, in-memory analytics and traditional data store (e.g, data warehouse/ data mart) to help enable speed and faster insight.

A big data store is a distributed file system and compute engine that can be used to facilitate data movement and processing. It contains what we call a big data core — a computer cluster with distributed data storage and computing power. Advancements in big data technologies have enabled big data store to function as a platform for additional types of computing, some of which (like query engines) can specifically support data interactivity. Traditionally, the big data core file system can use techniques such as replication and sharing (database partitioning that separates very large databases into smaller, faster, more easily managed parts) to accelerate and scale data storage. Additionally, these techniques can help strengthen processing capabilities. Newer additions enable more powerful use of the core memory as a high-speed data store, supporting improved data movement, processing, and interactivity. These improvements allow for in-memory computing on an existing computer cluster. Moreover, streaming technologies added to the core can enable real-time complex event processing, and in-memory analytics technologies support better data interactivity.

An in-memory analytics is a database management system that relies primarily on main memory for computer data storage instead of the disk storage mechanism used by traditional database management systems. In-memory databases are faster because the internal algorithms are simpler and execute fewer CPU instructions. Moreover, accessing data in memory eliminates the "seek time" involved in querying data on disk storage, thus providing speedier and more predictable performance. Because in-memory platform constrain the entire database and the applications to a single address space, they reduce the complexity of data management. Any data can be accessed within just microseconds.

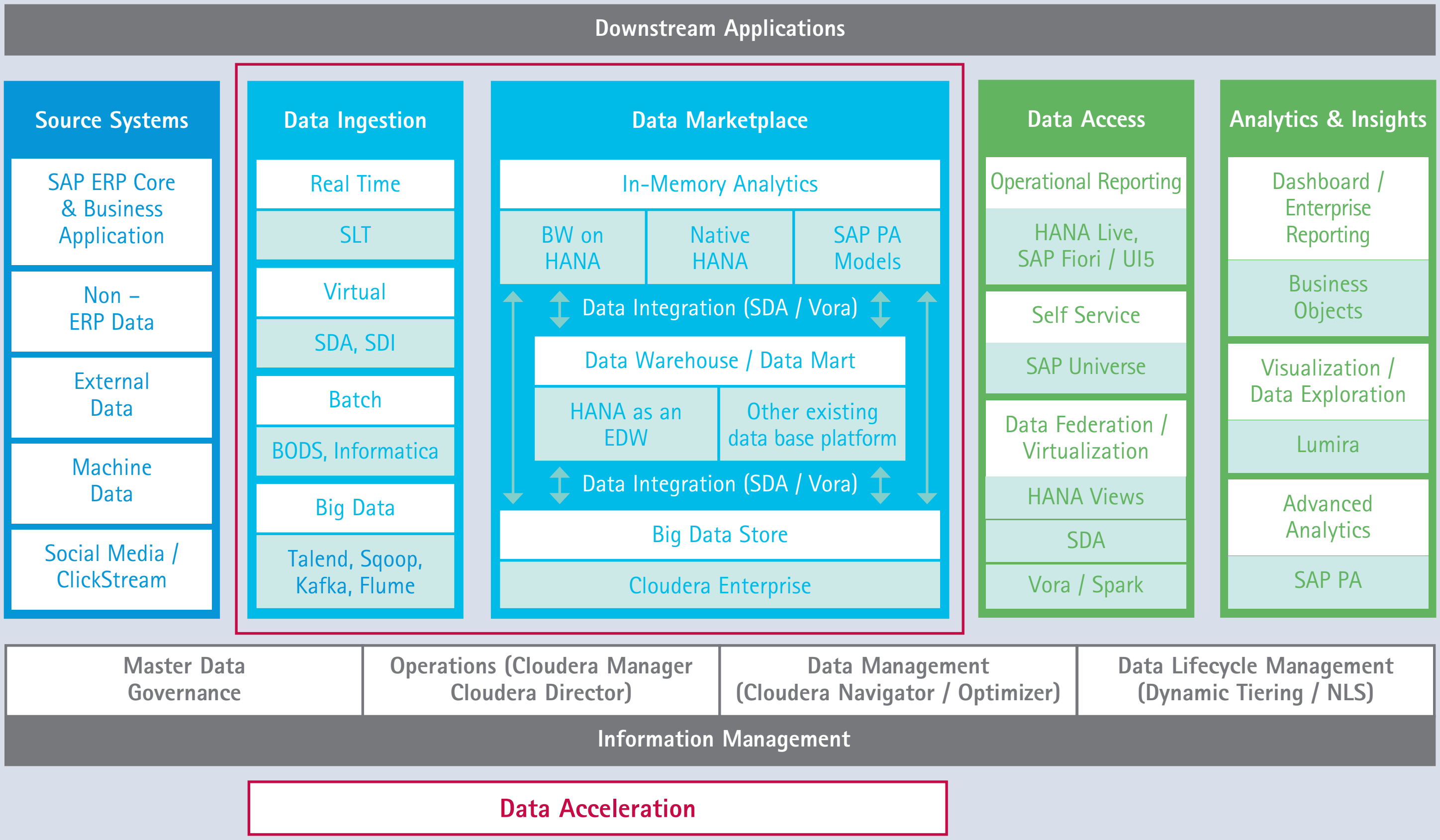
Combining components to create solutions for organizations running SAP

With an understanding that the majority of organizations running SAP will eventually adopt SAP's in-memory technology, SAP HANA, as one of their data platforms, our point of view is that they should capitalize on that investment and make it a key part of their modern data supply chain and compliment/integrate it with a big data platform such as Cloudera. The combination of these technologies will help enable organizations to conquer the three data related challenges while maximizing the return on investments.

To highlight how these integrated data platforms (SAP HANA and Cloudera) can help modernize the data supply chain and accelerate an SAP organization's insights, the following sections outline:

1. A reference architecture using SAP HANA and Cloudera for data acceleration (some illustrative use cases included)
2. How Cloudera Enterprise addresses the three data challenges for organizations with SAP HANA
3. Pragmatic next steps to help organizations start the journey to ROI for analytics

FIGURE 3: A reference architecture using SAP HANA and Cloudera for data acceleration



Downstream Applications
Source Systems
SAP ERP Core & Business Application
Non – ERP Data
External Data
Machine Data
Social Media / ClickStream
Data Ingestion
Real Time
SLT
Virtual
SDA, SDI
Batch
BODS, Informatica
Big Data
Talend, Sqoop, Kafka, Flume
Data Marketplace
In-Memory Analytics
BW on HANA
Native HANA
SAP PA Models
Data Integration (SDA / Vora)
Data Warehouse / Data Mart
HANA as an EDW
Other existing data base platform
Data Integration (SDA / Vora)
Big Data Store
Cloudera Enterprise
Data Access
Operational Reporting
HANA Live, SAP Fiori / UI5
Self Service
SAP Universe
Data Federation / Virtualization
HANA Views
SDA
Vora / Spark
Analytics & Insights
Dashboard / Enterprise Reporting
Business Objects
Visualization / Data Exploration
Lumira
Advanced Analytics
SAP PA
Master Data Governance
Operations (Cloudera Manager Cloudera Director)
Data Management (Cloudera Navigator / Optimizer)
Data Lifecycle Management (Dynamic Tiering / NLS)
Information Management
Data Acceleration

Cloudera Enterprise complements SAP architecture to address the data related challenges of movement, processing, and interactivity

Movement

Hadoop, by its nature, is able to handle data of all types and of all volume. Cloudera Enterprise, with the inclusion of ingestion components such as Apache Sqoop, Apache Flume, and Apache Kafka, is able to scale out and speed up the ingestion of data from all sources.

Apache Sqoop is a component of Cloudera Enterprise that enables the ingestion of structured data stored in relational databases. Sqoop allows bi-directional data movement between Cloudera Enterprise and virtually any relational database system and NoSQL systems. Apache Sqoop includes high performance connectors available for a number of enterprise data warehouse systems.

Apache Flume enables any streaming source such as a web or application server, network device, or operational system to load streaming data directly into Cloudera's data storage engines.

Apache Kafka is a distributed publish-subscribe messaging system that is designed to be fast, scalable, and durable. This open source project – licensed under the Apache license – has gained popularity within the Hadoop ecosystem, across multiple industries. Its key strength is the ability to make high volume data available as a real-time stream for consumption in systems with very different requirements — from batch systems like Hadoop, to real-time systems that require low-latency access, to stream processing engines like Apache Spark Streaming that transform the data streams as they arrive. Kafka's flexibility makes it ideal for a wide variety of use cases, from replacing traditional message brokers, to collecting user activity data, aggregating logs, operational application metrics and device instrumentation.

Through this collection of data ingestion components, Cloudera Enterprise addresses the data movement challenge and offers:

- Low TCO for data ingestion of a wide variety of data
- Scalable ingestion via a parallel processing framework
- Supports real-time and non-real-time types of applications and data sources
- Enables an enterprise data hub for immediate access by users

Processing

Cloudera Enterprise provides faster and scalable data storage engines such as Apache HBase and Kudu that serve data to analytic and reporting pipelines with low latency and high availability. Engines such as Apache Impala (incubating) provide SQL APIs to data stored in Cloudera Enterprise. For faster general data processing, Spark provides a simple API for data scientists to do sophisticated analytics.

Apache HBase is the high-performance, distributed data store built for Apache Hadoop that perform fast, random reads and writes to all data stored. It integrates with other components, like Apache Kafka or Apache Spark Streaming, to build complete end-to-end workflows all within the single platform.

Apache Impala (incubating) is an open source, analytic database that runs natively in Apache Hadoop. Impala combines all of the benefits of other native Hadoop frameworks, including flexibility, scalability, and cost-effectiveness, with the performance, usability, and SQL functionality necessary for traditional databases. Impala is an integrated part of Cloudera's enterprise data hub – sharing the same flexible file and data formats, metadata, security, and resource management as components such as MapReduce, Apache Hive, and Apache Pig. It also seamlessly integrates with popular third-party tools, such as SAP BusinessObjects, allowing enterprises to continue to leverage existing investments.

Apache Kudu is a columnar storage manager developed for the Hadoop platform. Kudu shares the common technical properties of Hadoop ecosystem applications: it runs on commodity hardware, is horizontally scalable, and supports highly available operation. Kudu works as the low-latency storage engine that serves data to the SAP ecosystem.

Interactivity

Cloudera has certified a number of reporting and visualization engines such as SAP Lumira and SAP Business Objects, which means that users can explore data in SAP HANA and Cloudera Enterprise in a seamless fashion and with low latency.

Data scientists can process data using Apache Spark's in memory data processing and machine learning capabilities to tap into integrated enterprise data and uncover game changing hypothesis. Spark is also leveraged by SAP HANA Vora to process OLAP and analytic pipelines on data stored in the Hadoop/SAP ecosystem.

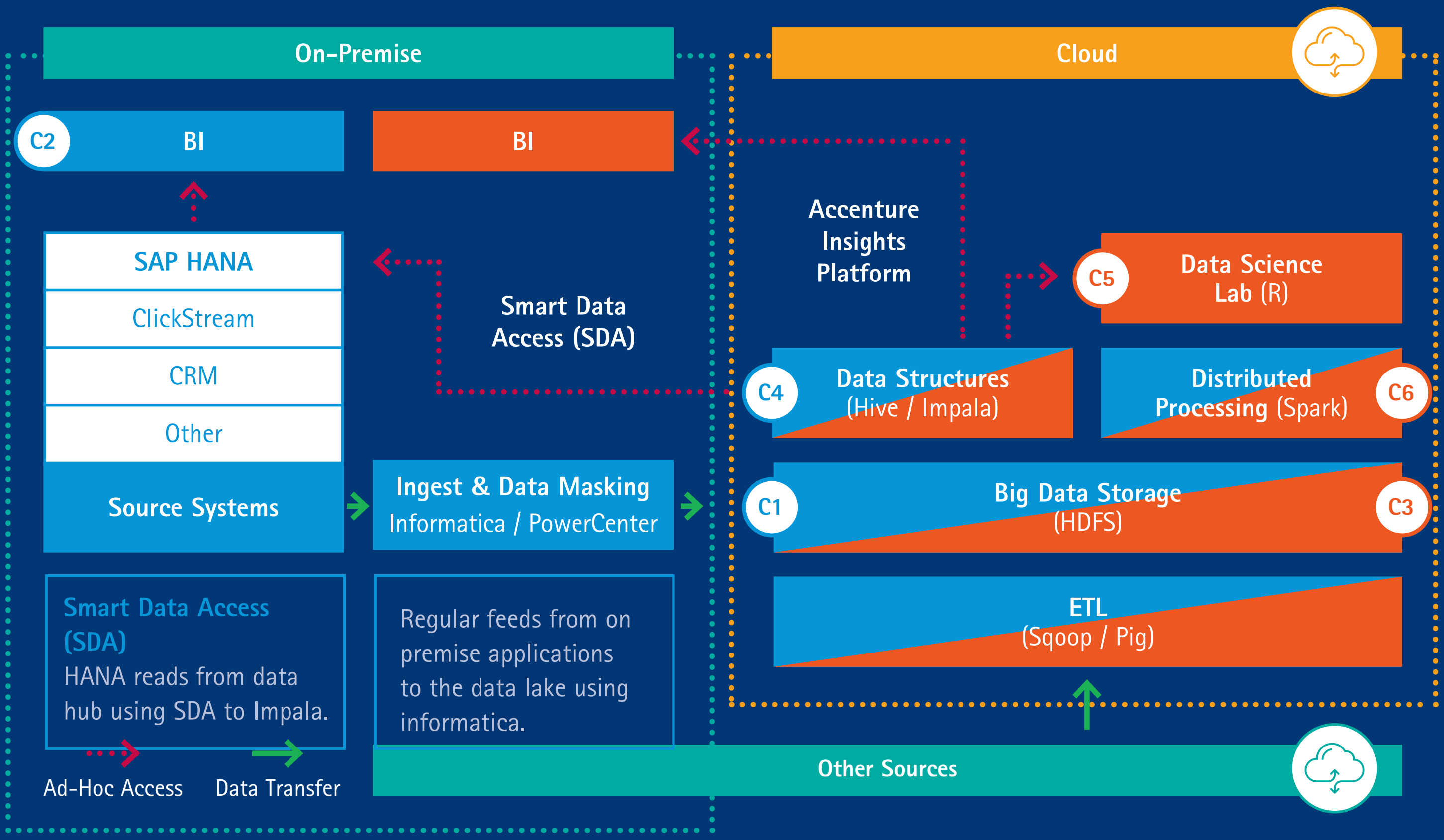
Apache Spark is the open standard for flexible in-memory data processing that enables batch, real-time, and advanced analytics on the Apache Hadoop platform. Via the One Platform initiative, Cloudera is committed to helping the ecosystem adopt Spark as a replacement for MapReduce in the Hadoop ecosystem as the default data execution engine for analytic workloads.

Cloudera's key capabilities for addressing the data interactivity challenge:

- Low latency engine for interactive queries: Apache Impala (incubating)
- Smart Data Access (SDA) – Data in Hadoop can be registered as tables in SAP HANA and queried as if they were tables in SAP HANA
- SAP HANA Vora is a engines (Running in a Cloudera Cluster) can access data in SAP HANA
- Most industry standard reporting and visualization engines such as SAP Lumira and SAP Business Objects have been certified to operate with Cloudera Enterprise can access data stored in the SAP / Cloudera ecosystem.

FIGURE 4: Reference architecture applied: Example 1 – Customer

Sample big data component and integration architecture for optimized concurrent ad-hoc reporting using certified SAP HANA integration via Apache Impala (incubating) and Informatica.



Smart Data Access (SDA)
HANA reads from data hub using SDA to Impala.

Regular feeds from on premise applications to the data lake using informatica.

	Enabling Capacity
C1	Analytical Data Model
C2	Connected Reporting
C3	Batch Data Processing
C4	Batch Data Provisioning
C5	Statistical Modeling
C6	Batch Model Execution

Accenture Insights Platform
Provides Analytics and Big Data fully cloud based. SW licenses and hosting are included in a regular monthly fee. Systems have full scaling flexibility.

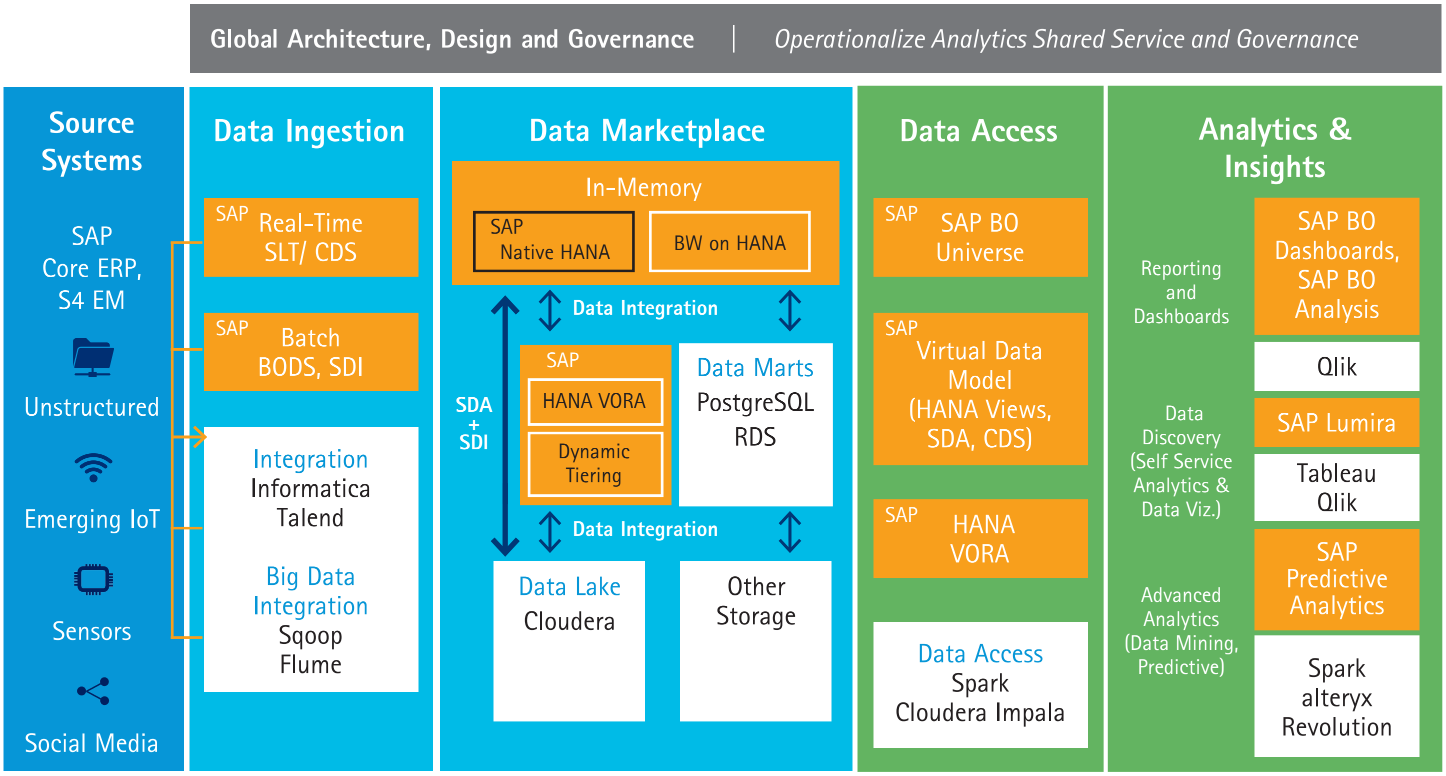
Data Science Lab (R)
Scalable data science platform based on Revolution R. R reads data directly from the lake to minimize data provisioning durations and maximize analysis time.

Big Data Storage (HDFS)
Cloudera based Hadoop Data Lake. Authorizations and Authentication are managed in RecordService (Beta version) and Kerberos.

Lab	Factory
Data Scientists	Business User
Light Governance	Full Strict Governance
Speed to Insights, Flexibility, Capacity	Stability, Scalability

FIGURE 5: Hybrid Data and Analytics Platform based on SAP HANA and Accenture Insights Platform (AIP) leveraging Cloudera

Combine assets and accelerators based on HANA On Premise and Accenture Insight Platform (AIP) as-a-Service options to meet unique needs and help drive outcomes. The Accenture Insights Platform is a comprehensive and scalable solution that allows organizations to get actionable insights and business outcomes quickly with a competitive, flexible, consumption-based commercial model.



Digital Transformation driving new Data Sources

Data Ingestion layer that allows various data elements – type & latency to flow into the storage environment

Data Hub allows to store & transform hot / warm / cold data with minimal data movement

Data Access providing the semantic layer and promotes re-use of data and models across domains

Analytics & Insights Provide tools to make it easier to get insights from data ranging from simple reports to complex statistical models

SAP (On-Premise, aaS)

Based on Accenture Insight Platform

Pragmatic next steps to help you on your journey

Many companies, already reeling from the impacts of technology and the changes they need to make in response, find themselves temporarily overwhelmed — some even paralyzed as they absorb the magnitude of the tasks ahead. That's understandable.[5]

To help organizations realize the value of accelerated data in a modern supply chain, we recommend the following three step best practice:

1

PROJECT

We start with a short data discovery Proof of Concept, demonstrating how to solve a strategic business problem with client data, leveraging an intuitive visualization to deliver findings.

2

WORKSHOP

We all work with clients to enable the organization to build a roadmap, using facilitation methodology focused on outcomes, leveraging the value in data that has been unlocked by analytics in a collaborative environment.

3

FACILITY

We embed a technology-enhanced facility to render data insights more accessible to decision makers at all levels of the organization, enabling data discovery and non-linear storytelling within workshops and meetings.

About Accenture & Cloudera Alliance

About Cloudera

Cloudera delivers the modern data management and analytics platform built on Apache Hadoop and the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest and most secure data platform available for the modern world. Our customers efficiently capture, store, process and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly and at lower cost than has been possible before. To ensure our customers are successful, we offer comprehensive support, training and professional services.

Accenture & Cloudera Collaboration – Enabling High Performance

Accenture and Cloudera help companies cohesively mobilize, access and manage big data across the enterprise to help drive actionable insights, make more informed decisions and achieve better business outcomes.

We help deliver end-to-end big data solutions at enterprise scale so you can use more of your data, reduce the need to limit or move large datasets and also centralize information security, metadata, management and governance.

We help build a technology ecosystem to drive business results from big data to enable data access and discovery, delivering multi-genre analytics in a single platform, and protecting existing technology and skillset investments.

We help shorten the analytics lifecycle by moving from big data to big outcomes at pace and speeding data access, discovery and time-to-insights.

This collaboration combines Accenture's industry and digital experience, its analytics skills and the focus on analytics-driven outcomes through Accenture Analytics, with Cloudera's enterprise data hub built on Hadoop. Our alliance strengthens Accenture's commitment to bringing emerging big data and analytics technologies to our clients and helping them build effective and outcome-driven data management platforms.[6]

Reference Material

[1] **Accenture Analytics Journey to ROI**
www.accenture.com/us-en/analytics-cross-industry-journey

[2] **Accenture Technology Vision 2016**
www.accenture.com/us-en/insight-technology-trends-2016.aspx

[3] **High velocity enterprises are changing the game by reimagining ERP**
www.accenture.com/us-en/insight-erp-sap.aspx

[4] **Accenture Connected Analytics Experience**
www.accenture.com/us-en/service-accen

[5] **Data Acceleration: Architecture for the Modern Data Supply Chain**
www.accenture.com/us-en/insight-data-acceleration-modern-data-supply-chain.aspx

[6] **Accenture/ Cloudera partnership**
www.accenture.com/us-en/service-accenture-cloudera-data-center-business-transformation

Who To Contact

Thad Gustafson
MD, Technology Innovation & Ecosystems
thad.gustafson@accenture.com

Amiya Chand
SM, Digital Analytics Accenture
amiya.a.chand@accenture.com

Bob Gressens
Director, Accenture Global Alliance, Cloudera
bgressens@cloudera.com

About Accenture

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions—underpinned by the world's largest delivery network—Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With more than 375,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at www.accenture.com.