Cloud Customer Architecture for Big Data and Analytics V2.0

Executive Overview

Big data analytics (BDA) and cloud are a top priority for most CIOs. Harnessing the value and power of data and cloud can give your company a competitive advantage, spark new innovations, and increase revenues. As cloud computing and big data technologies converge, they offer a cost-effective delivery model for cloud-based analytics.

As a delivery model, cloud computing has the potential to make your business more agile and productive, enabling greater efficiencies and reducing costs. Many companies are experimenting and iterating with different cloud configurations as a way to understand and refine requirements for their big data analytics solutions without upfront capital investment.

Both technologies continue to evolve. Organizations know how to store big data—they now need to know how to derive meaningful analytics from the data to respond to real business needs. As cloud computing matures, more enterprises are building efficient and agile cloud environments, and cloud providers continue to expand service offerings.

This paper describes a well-tested reference architecture for Big Data and Analytics in a hybrid cloud environment. In addition, you will:

- Discover business reasons for organizations to adopt cloud for their analytics needs.
- Receive an architectural overview of an analytics solution in a cloud environment with a description of the capabilities offered by cloud providers.
- Learn about the architectural components of the solution.
- Review example architectural scenarios that might be similar to your operating environment.

Business drivers

There are many reasons businesses should adopt big data and analytics capabilities in their organization and use cloud computing to enable those capabilities. Specific business drivers include:

- **Low up-front cost**: The cloud computing delivery model lets you set up new analytics infrastructure quickly and test new scenarios without incurring significant up-front expenses. If the exploration and analytics do not provide expected business value, you can quickly tear down the analytics environments.
- **Speed and agility:** The cloud delivery model enables clients to rapidly establish an analytics infrastructure without the usual lead times of infrastructure ordering, provisioning, and the like.

The agility provided by the cloud lets you quickly scale your analytics infrastructure up and down as data volumes change.

- **Ability to keep up with changing analytics capabilities:** Data platforms and analytics capabilities are rapidly changing, creating an almost-constant need for new technology to keep up with the latest trends. Cloud providers typically keep their service catalogs updated as new technologies evolve, providing the best model for consuming evolving analytics capabilities. Similarly, as data volumes grow, new technologies are emerging that deliver scale-out data transfer, enabling efficient, large-scale workflows for ingesting, sharing, collaborating, and exchanging big data. Cloud delivery models keep the cost of change low, especially as you commission and decommission new technology.
- **Advancement in cloud security:** Security within a cloud environment has always been a prime concern, especially for organizations in highly regulated industries. Cloud providers now offer ways to enforce security at multiple layers (such as data, network, authentication and authorization, privileged user monitoring) and processes to demonstrate compliance to a number of industry standards such as PCI, HIPAA, etc.
- **Reduced barrier to entry and new business models:** With the advent of the cloud, startups and small businesses now have the means to enter industries that previously demanded large up-front investments.
- **Innovation:** The convergence of big data, analytics and cloud is fueling innovation. With the low costs, speed, agility, and security that the cloud offers, companies have more time and money to experiment and bring to life the latest innovative technology.

A set of typical big data and analytics use cases for various industries are included in the Appendix. The new reference architecture proposed in this paper can be used to create cloud-based big data and analytics solutions for solving these business scenarios and help drive business success.

Architecture Overview

The big data and analytics cloud architecture guidance provided by this paper can help enterprises understand proven architecture patterns that have been deployed in numerous successful enterprise projects.

Cloud deployments offer a choice of private, public and hybrid architectures. Private cloud employs in-house data and processing components running behind corporate firewalls and dedicated cloud. Public cloud offers services over the Internet with data and computing resources available on publicly assessable servers. Hybrid environments have a mixture of components running as both in-house, dedicated cloud and public cloud services with data flowing between them.

There are also choices in the levels of services that a cloud provider can offer for an analytics solution:

- Basic data platform infrastructure service – such as Hadoop as a service – providing pre-installed and managed infrastructures. With this level of service, you are responsible for loading, governing and managing the data and analytics for the analytics solution.

- A governed data management service – such as a data lake service [1] – providing data management, catalog services, analytics development, security and information governance services on top of one or more data platforms. With this level of service, you are responsible for defining the policies for how data is managed and for connecting data sources to the cloud solution. The data owners have direct control of how their data is loaded, secured and used. Consumers of data are able to use the catalog to locate the data they want, request access and make use of the data through self-service interfaces.
- An insight and data service – such as a customer analytics service. With this level of service, you are responsible for connecting data sources to the cloud analytics solution. The cloud analytics solution then provides APIs to access combinations of your data and additional data sources, both proprietary to the solution and public open data, along with analytical insight generated from this data.

It is important to have this choice of cloud deployment because data and processing location is one of the first architectural decisions for an analytics cloud project. This allows both flexibility in operating models and the optimal placement of both data and analytics workloads on the available processing platforms. Legal and regulatory requirements may also impact where data can be located since many countries have data sovereignty laws that prevent data about individuals, finances and certain types of intellectual property from moving across country borders.

The choice of cloud architectures allows compute components to be moved near data to optimize performance when data volume and/or bandwidth limitations result in bottlenecks for remote data and data movement.

Figure 1 illustrates a simplified enterprise cloud architecture for a big data and analytics environment. The architecture has three network zones: public network, provider cloud, and enterprise network.

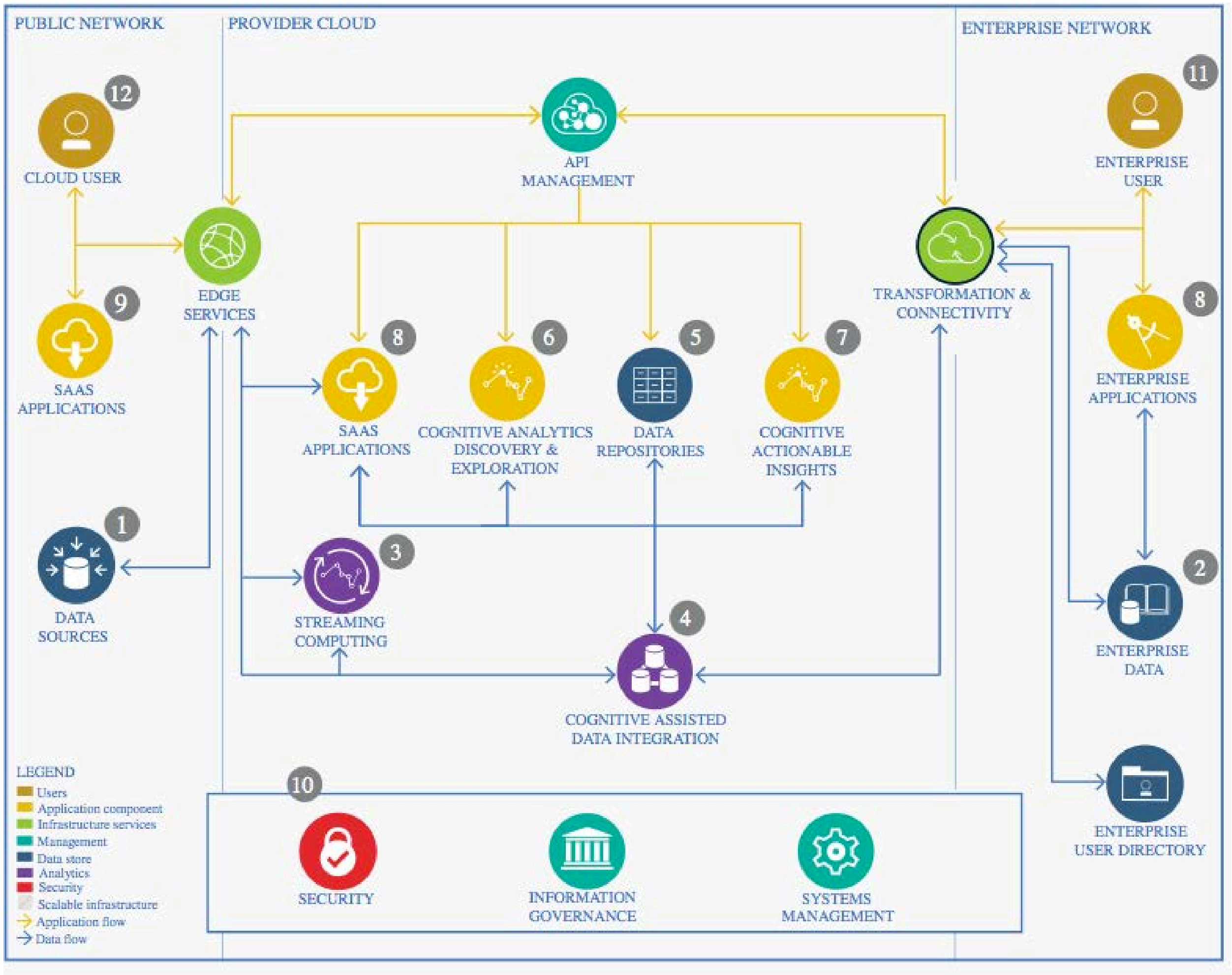


Figure 1: Big Data & Analytics Solutions in the Cloud

This big data and analytics architecture in a cloud environment has many similarities to a data lake deployment in a data center. Data is collected from structured and non-structured data sources. It is staged and transformed by data integration and stream computing engines and stored in different data repositories. The data is transformed, augmented with analytical insight, correlated and summarized as it is copied and moved through this processing chain and selected data is made available to consumers through APIs. Cognitive technologies such as machine learning, deep learning and natural language processing can be leveraged to semi-automate data ingestion, data integration, analytics discovery and exploration, and actionable analytics.

Information governance, security and system management encompass each processing phase to ensure regulation and policies for all data are defined, enabled and reinforced across the system. Compliance is tracked to ensure controls are delivering expected results. Cloud Security covers all elements including generated data and analytics, as well as the underlying security of the cloud platform.

The users of the analytics solution are broadly classified in two ways: enterprise and third party, or cloud users. Enterprise users access resources on premises or via a secure Virtual Private Network (VPN). Data is available directly and through applications that provide reports and analytics. Transformation and connectivity gateways prepare information for use by enterprise applications as well as use on different devices, including mobile, web browsers and desktop systems. Third party users gain access to the provider cloud or the enterprise network via edge services that secure access to users with proper credentials. Access to specific data sets and analytics are then typically further restricted as dictated by corporate policy and controlled by the appropriate data owners.

The above architectural approach supports the entire lifecycle of analytics, enabling the deployment of production analytics, as well as a data lake type of architecture that serves as a DevOps environment for data, collaboration and analytics. This is achieved by the addition of common metadata and semantic definitions to the descriptions of enterprise data repositories that are stored in a catalog. These catalog entries are augmented with governance classifications, rules and policies that are used by the processing engines in the data lake to automate the management of the data as it flows in, out and through the data lake. Additional data repositories provide sandboxes of selected data for analytical models and places for users to store their own data. Together these data repositories provide the data for the development of new analytics models or enhancements of existing models.

Component Model

The following sections provide a summary of each of the major components.

Public Network Components

The public network contains elements that exist in the Internet: cloud users, SaaS applications, data sources and edge services.

Cloud User

A cloud user is a person that connects to the analytics cloud solution via the Internet. This person may be uploading new data, searching and retrieving data, providing feedback on the data, requesting new analytics or data, or running existing analytics.

Users of the analytics solution in the cloud may be performing different roles, including:

Knowledge Worker & Citizen Analyst
Data Scientist
Application Developer
Data Engineer
Chief Data Officer (CDO)
CLOUD USER

- **Knowledge worker and citizen analyst:** Business analysts who are subject matter experts (SMEs) in the actual business are demanding a highly agile self-service model that allows them to find information relevant to a topic they are analyzing and discover insight in that data.
- **Data scientist:** Typically, data scientists are trained in a quantitative discipline such as statistics, operations research, machine learning, econometrics or an equivalent field. They have a deep understanding of the mathematical and computational methods that can be applied to data to derive insight for the business process. Data scientists discover insight in a rich set of data, and this discovery includes preparing, cleansing and enriching data from different sources, both internally and externally.

- **Application developer:** These individuals incorporate actual analytics algorithms into an application, often in the shape of scoring functions, that will be integrated with a business process run on a production-level system and data model.

- **Data engineer:** As a traditional IT persona who manages the data, data engineers build physical and logical models. They are also responsible for the integration tasks to bring capabilities defined by data scientists, application developers and Chief Data Officers (CDOs) into production-level business processes.

- **Chief Data Officer (CDO):** As executive-level owners of an organization's data, CDOs define the logical business-object models and governance rules including data access policies. They are ultimately responsible for the quality of the data.

All the different personas have the following common characteristics:

- They want self-service; they often take the do-it-yourself approach, with the ability to create sandboxes to test out new hypotheses and move the actionable insight to production.

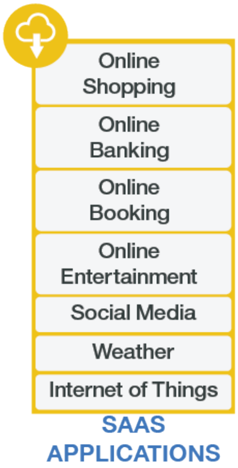
- They want access to the right data to accomplish the analytical task (this might include large volumes of data), no matter where the data is stored, with a good understanding of the data quality and lineage.

- They often need access to many different tools and capabilities, many of which are open-source based, and can be on-demand (based on workload and scaling).

- Lastly, they need to collaborate with each other. One way the cloud providers enable this is by building a knowledge base using Graph technology and allowing the user to see the correlation of the users and data, i.e., who used what data and how the data was used, thus identifying the right people with whom the user should collaborate.

SaaS Applications

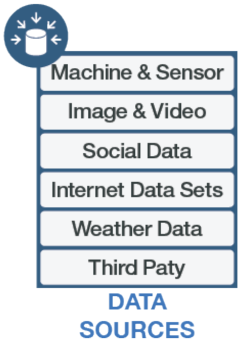
Increasingly, organizations are making use of applications offered as a cloud service. This type of cloud service is called software-as-a service, or SaaS. The SaaS applications on the public network are mostly mobile apps and web applications to engage with customers, for example, online-banking, online-shopping, online-booking for travel, IoT applications such as connected cars and smart house, weather and social media, etc.



Data Sources

Public data sources contain external sources of data for the data analytics solutions that flow from data providers through the Internet.

There can be a number of different information sources in a typical big data system, some of which enterprises are just beginning to include in their data analytics solutions. High velocity, volume, variety, and data inconsistency often kept many types of data from being used extensively. Big data tools have enabled organizations to use this data; however, these typically run on-premises and can require substantial upfront investment. Cloud computing helps mitigate that investment and the associated risk by providing big data tools via a pay-per-use model. It also allows edge analytics services meaning that the data analytics is applied where the data is generated producing real time analytics at the edge. Data sources include:



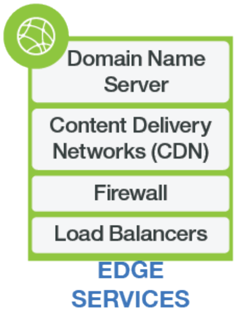
- **Machine & Sensor:** Data generated by devices, sensors, networks and related automated systems including Internet of Things (IoT).
- **Image & Video:** Data capturing any form of media (pictures, videos, etc.) which can be annotated with tags, keywords and other metadata.
- **Social:** Data for information, messages and pictures/videos created in virtual communities and networks.
- **Internet Data Sets:** Data stored on websites, mobile devices and other internet-connected systems.
- **Weather Data**
- **Third Party**: Data used to augment and enhance existing data with new attributes like demographics, geospatial or CRM.

Edge Services

Edge services include services that allow data to flow safely from the Internet into the data analytics processing system hosted on either the cloud provider or in the enterprise.

When the data or user requests come from the external Internet, the flow may come through edge services including Domain Name System (DNS) servers, Content Delivery Networks (CDNs), firewalls, and load balancers before entering the cloud provider's data integration or data streaming entry points.

Edge services also allow users to communicate safely with the analytical system and enterprise applications. These include:



- **Domain Name System Server:** Resolves the URL for a particular web resource to the TCP-IP address of the system or service, which can deliver that resource.
- **Content Delivery Networks (CDN):** Provide geographically distributed systems of servers deployed to minimize the response time for serving resources to geographically distributed users, ensuring that content is highly available and provided to users with minimum latency. Which servers are engaged will depend on server proximity to the user and where the content is stored or cached. CDNs are typically for user flows and not data source flows.
- **Firewall:** Controls communication access to or from a system, only permitting traffic that meets a set of policies to proceed and blocking any traffic that does not meet the policies.
- **Load Balancers:** Provide local or global distribution of network or application traffic across many resources (such as computers, processors, storage, or network links) to maximize throughput, minimize response time, increase capacity, and increase reliability of applications. Load balancers should be highly available without a single point of failure. Load balancers are sometimes integrated

as part of the provider cloud analytical system components like stream processing, data integration, and repositories.

Provider Cloud Components

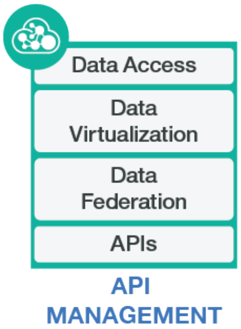
The provider cloud represents the cloud-based analytics solution. It hosts components to prepare data for analytics, store data, run analytical systems, and process the results of those systems. Provider cloud elements include:

- API management
- Streaming computing
- Cognitive assisted data integration
- Data repositories
- Cognitive analytics discovery and exploration
- Cognitive actionable analytics
- SaaS applications
- Transformation and connectivity

Data Access and API Management

The overall purpose of the Data Access component is to express the various capabilities needed to interact with the Data Repositories component. The capabilities serve the access needs of data scientists, business analytics, developers, and others that need access to valuable data.

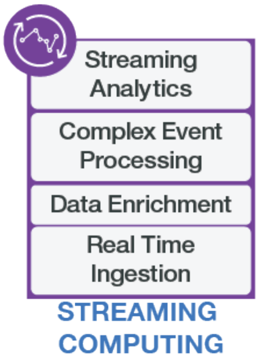
It includes the following types of services:



- **Data Access:** Capabilities to push and pull data in and out of the data lake repositories by the users.
- **Data Virtualization:** Describes any approach to data management that allows a user or application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted or where it is physically located.
- **Data Federation:** Often used in conjunction with data virtualization, but is the specific ability to have a single query interface to various different data repositories - allowing for a single, more usable, view of the data to business analysts and report writers.
- **Open APIs:** With the ever changing technology landscape there is a continued need to allow for multiple types of open APIs access channels to the data lake repository. This capability represents that fluid list of technologies and solutions.

Streaming Computing

Stream processing systems can ingest and process large volumes of highly dynamic, time-sensitive continuous data streams from a variety of inputs, such as sensor-based monitoring devices, messaging systems, and financial market feeds. The “store-and-pull” model of traditional data-processing environments is not suitable for this class of low-latency or real-time streaming application where data needs to be processed on the fly as it arrives. Capabilities include:



- **Streaming Analytics**: Applying analytic processing and decision making to in-memory and transient data with minimal latency.
- **Complex Event Processing (CEP)**: Event processing that combines data from multiple sources to identify meaningful events or patterns (such as opportunities or threats) and respond to them as quickly as possible.
- **Data Enrichment:** Combining in-motion data with the historical data from the data lake for real-time or near real-time analytics.
- **Real Time Ingestion**: This capability can be leveraged to ingest and prepare structured or unstructured data arriving from source systems in real-time or near-real-time. The capability deals with transactional data transmitted via a message hub or enterprise service bus, as well as streaming data such as sensor data or video feeds.

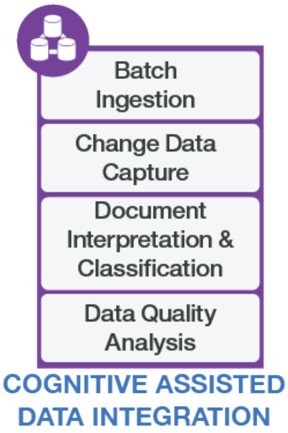
Cloud services allow streaming computing to be adapted as data volume and velocity changes. Peaks in demand can be accommodated by adding virtual memory, processors and storage. The option to add dedicated hardware can also help with specialized processing needs.

Cognitive Assisted Data Integration

The Cognitive Assisted Data Integration component focuses on the processes and environments that deal with the capture, qualification, processing, and movement of data in order to prepare it for storage in the Analytical Data Lake repositories, which is subsequently shared with the Discovery & Exploration and Actionable Insights components, via the Data Access component. The Data Ingestion & Integration component may process data in scheduled batch intervals or in near real-time/"just-in-time" intervals, depending on the nature of the data and the business purpose for its use. Various cognitive technologies such as machine learning and natural langurage processing can be leveraged to semi-automate the data ingestion and integration process.

Data to be integrated can come from public network data sources, provider cloud SaaS applications, enterprise data sources, or streaming computing results. The results from data integration can be passed to data repositories for analytical processing, or passed to enterprise data for storage or feeding into enterprise applications.

Capabilities required for data integration include:


Batch Ingestion
Change Data Capture
Document Interpretation & Classification
Data Quality Analysis
COGNITIVE ASSISTED DATA INTEGRATION

- **Batch Ingestion**: This capability can be leveraged to ingest and prepare structured or unstructured data in batch mode, either on-demand or at scheduled intervals. The capability includes standard Extract, Transform and Load (ETL) and Extract, Load and Transform (ELT) paradigms, in addition to manual data preparation and movement.
- **Change Data Capture:** This capability can be leveraged for replicating data in near real time to support data migrations, application consolidation, data synchronization, dynamic warehousing, master data management (MDM), business analytics and data quality processes.
- **Document Interpretation & Classification:** This capability streamlines the capture, recognition, and classification of business documents, to

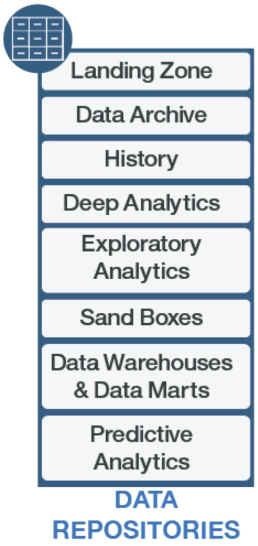
quickly and accurately extract important information from those documents for use by business users and in applications.

- **Data Quality Analysis:** This capability allows data quality rules (stored and maintained in the Information Management & Governance component) to be applied to data during ingestion and transformation, and for quality measures to be stored as metadata associated with the data sets in the analytics environment.

Data Repositories

The Data Repositories component is a set of secure data repositories that allows data to be stored for subsequent consumption by analytics tools and users. These repositories form the heart of the analytics environment. The repositories within this component may vary from a single Hadoop repository or Enterprise Data Warehouse, to multiple repositories used for different purposes by different analytical tools. Note that operational and transactional data stores (such as OLTP, ECM, etc.) are not included in this component. Instead they form part of the Data Sources component.

Types of data repositories include:



- **Landing Zone & Data Archive** is typically an initial location for data ingested from source systems and where raw data is persisted for archive purposes. Data in the Landing Zone may be of varying types and formats. It may or may not be modeled or structured, and its quality may or may not be understood.
- **History** is a repository that stores state-change histories, log data, etc. Such repositories are typically optimized for write operations, and are used for data that will not normally be accessed often or with real-time response requirements.
- **Deep & Exploratory Analytics** is the application of sophisticated data mining and analysis techniques to yield insights from large, typically heterogeneous data sets. Deep & Exploratory Analytics repositories are typically optimized for low-cost storage of very large data volumes, without the need for real-time response. These repositories are used to store shared, heterogeneous data for use by data scientists.
- **Sand Boxes** are repositories used by individual data scientists or groups who need a temporary data repository to experiment and do quick analyses. Sand boxes are provisioned, populated, deleted and re-deployed more often than other data repositories.
- **Data Warehouses & Data Marts** are analogous to the traditional Enterprise Data Warehouse and Data Marts, and are used to store data which will be read and analyzed frequently, with interactive real-time response requirements.
- **Predictive Analytics** are used to support operationalized predictive or prescriptive analytics use cases such as *next best action* scenarios, where real-time response is a requirement to support interactive workloads.

Cognitive Analytics Discovery & Exploration

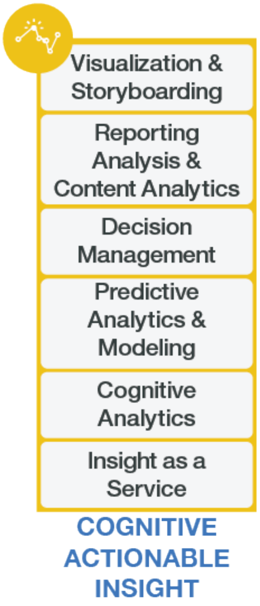
Data Science
Search
COGNITIVE ANALYTICS
DISCOVERY &
EXPLORATION

The overall purpose of the Discovery & Exploration component is focused around enabling a new (and old) breed of data customer. Data scientists, business analysts, data engineers, and application developers are required to find the hidden gems in the data quickly. Thus, they need the ability to collaborate about and easily interact with even the most complicated data repositories via new and emerging data science techniques. They also need the ability to semantically search across both structured and unstructured content to get a complete view of the data ontology.

- **Data Science:** This capability symbolizes a powerful new approach to making discoveries. By bringing together the areas of statistics, machine learning, deep learning, computer science and visualization, data science can bring insight and structure to the digital era. Data scientists have brought this ability to the forefront, but it is the presence of collaboration and tools that will bring data science to the masses.
- **Search and survey/shopping for data:** This capability provides the approach to use natural language processing techniques to search, learn, and discover things about desperate data. Think about having a single search and automated API facility to quickly understand important information from both structured and unstructured content.

Cognitive Actionable Insight

The overall purpose of the Actionable Insight component is to analyze data from a variety of sources in a cohesive manner and derive insight that is meaningful and actionable for the business domain. A variety of techniques are used to derive this valuable insight: visualization of the data, intelligent querying of multi-dimensional data, utilization of statistical models, data mining, content analytics, optimization and cognition. Cognitive technologies can be applied here to automatically select the right analytics model, to interact with users in a more human friendly way.



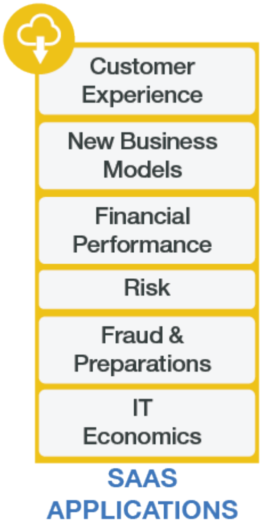
- **Visualization & Storyboarding:** Visualization capabilities help users in analyzing and reasoning about data by means of visual representation of the data. It makes complex data more accessible, understandable and usable. Storyboarding capability enables the user to organize a series of visualizations in sequence to effectively communicate an idea as a meaningful story.
- **Reporting, Analysis & Content Analytics:** This capability refers to the use of business intelligence techniques and solutions to analyze both structured content and unstructured content. These solutions answer predefined business questions and utilize high end visualizations to represent results in tabular view, graphs, charts, scorecards, and assembled dashboards with key performance indicators (KPI). Further, the content analytics capability augments the above mentioned capability and enables business analysis of unstructured content, such as text, images and video by utilizing analytical

methods and techniques.

- **Decision Management:** This capability improves the decision making process within organizations by generating actionable insight by using all available information to increase the precision, consistency and agility of decisions by taking known risks and constraints into consideration. Decision management makes use of business rules, statistical and optimization methods, and predictive analytics to automate the decision making process.
- **Predictive Analytics & Modeling:** Predictive analytics brings together advanced analytics capabilities spanning ad-hoc statistical analysis, predictive modeling, data mining, text analytics, entity analytics, optimization, real-time scoring, machine learning and more.
- **Cognitive Analytics:** This capability simulates human thought processes in a computerized model. It involves self-learning ability that use data mining, pattern recognition and natural language processing to mimic the way the human brain works. Actionable insight is derived as a result of the cognitive processing on the underlying domain data.
- **Insight as a Service:** The Insight as a service capability is the collection of accessible data domains that are readily available to utilize in advanced analytics applications or to push data back into data lake repositories.

SaaS Applications

Organizations are increasingly using SaaS applications on the provider network for various enterprise functions.



- **Customer Experience:** Customer-facing cloud systems can be a primary system of engagement that drives new business and helps service existing clients with lower initial cost.
- **New Business Models:** Alternative business models that focus on low cost, fast response and great interactions are all examples of opportunities driven by cloud solutions.
- **Financial Performance:** The office of finance should become more efficient as data is consolidated and reported faster and easier than in the past.
- **Risk:** Having more data available across a wider domain means that risk analytics are more effective. Elastic resource management means more processing power is available in times of heightened threat.
- **Fraud & Preparations:** Cloud solutions can provide faster access to more data allowing for more accurate analytics that flag suspicious activity and offer remediation in a timely manner.
- **IT Economics:** IT operations are streamlined as capital expenditures are reduced while performance and features are improved by cloud deployments.

Transformation and Connectivity

The transformation and connectivity component enables secure connections to enterprise systems with the ability to filter, aggregate, modify, or reformat data as needed. Data transformation is often required when data doesn't fit enterprise applications. Key capabilities include:

- **Enterprise Security Connectivity** monitors usage and secures results as information is transferred to and from the cloud provider services domain into the enterprise network to enterprise applications and enterprise data.
- **Transformations** transform data between analytical systems and enterprise systems. Data is improved and augmented as it moves through the processing chain.
- **Enterprise Data Connectivity** enables analytics system components to connect securely to enterprise data.

Enterprise Security Connectivity
Transformations
Enterprise Data Connectivity
TRANSFORMATION & CONNECTIVITY

Enterprise Network

The enterprise network is where the on-premises systems and users are located.

- Enterprise Users
- Enterprise Applications
- Enterprise Data
- Enterprise User Directory

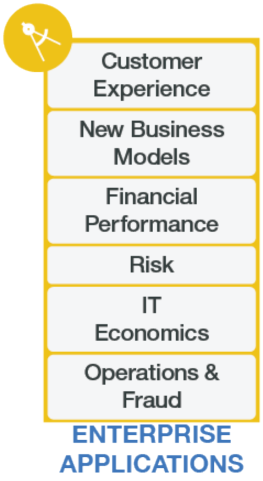
Enterprise Users

Enterprise users are individuals that act as both cloud users discussed above, and also as specialized users that connect to the analytics cloud solution through the organization's internal network. Enterprise users can also be administrative users, setting up the analytical processing system and monitoring solution performance and availability.

Enterprise Applications

Enterprise applications are key data sources for an analytics solution. They may also be a destination for new insight, or may act as a deployment platform for real-time analytic models developed in the data lake and can also provide information for SaaS applications.

See the descriptions of each subcomponent in the *SaaS Applications* section for the provider cloud (above).

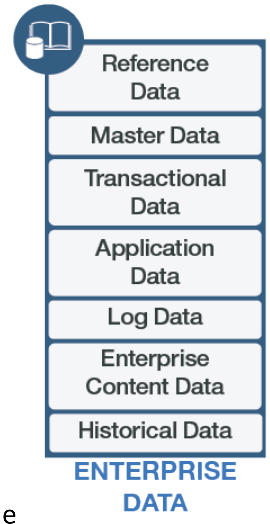


Enterprise Data

Within enterprise networks, enterprises typically host a number of applications that deliver critical business solutions along with supporting infrastructure like data storage. Such applications are key sources of data that can be extracted and integrated with services provided by the analytics cloud solution.

Enterprise data includes metadata about the data as well as systems of record for enterprise applications. Enterprise data may flow directly to data integration or the data repositories providing a feedback loop in the analytical system.

Enterprise data includes:



- **Reference Data:** This data provides authoritative lists of valid values and other types of look-up data (such as country codes and zip codes).
- **Master Data:** Master data provides selective attributes about key entities. These could be customers, products, assets, employees, or accounts. Typically, the data in a master data repository has been improved, augmented, and de-duplicated so it can be considered as an authoritative source of data. These repositories can be updated with the output of analytics to assist with subsequent data transformation, enrichment, and correlation. They can host analytics and feed other analytics models when they execute.
- **Transactional Data:** Data about or from business interactions. This data describes how the business operates.
- **Application Data:** Application data can come from applications running in the enterprise. It is a mixture of master, reference, transactional, and historical data blended together to support the operation of the application.
- **Log Data:** Data aggregated from log files for enterprise applications, systems, infrastructure, security, governance, and the like. Log data includes audit logs, website clickstream data, and error logs from processes.
- **Enterprise Content Data:** This is enterprise electronically stored information (i.e., word processing documents, emails, images, voice recording, or media data) that is managed in an enterprise content management solution. Enterprise content data is enriched with tags that describe its origin, the processes it has been generated from, and other contextual information.
- **Historical Data:** Data from past analytics and enterprise applications and systems that is saved based on business needs.
- **Archived Data:** Data from past analytics and enterprise applications and systems that have been archived due to legal reasons following corporate data retention policies.

Enterprise User Directory

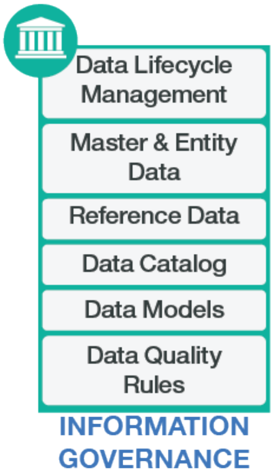
The enterprise user directory contains the user profiles for both the cloud users and the enterprise users. A user profile provides a login account and lists the resources (data sets, APIs, and other services) that the individual is authorized to access. The security services and edge services use this to drive access to the enterprise network, enterprise services, or enterprise-specific cloud provider services.

Cross All Environment Components

- Information Governance
- Security
- System Management

Information Governance

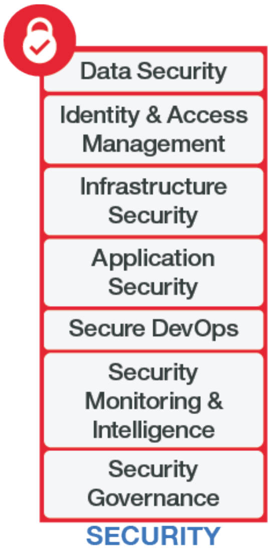
The Information Management and Governance components help you build confidence in your data by maintaining a trusted, accurate view of critical business data, providing a standardized approach to discovering your IT assets, and defining a common business language. The result is better, faster decision making that leads to operational efficiency and competitive advantage.



- **Data Lifecycle Management:** A policy-based approach to managing the flow of data throughout its life cycle. Provides records management, electronic discovery, compliance, storage optimization, data migration, data archive, and data disposal.
- **Master & Entity Data:** Provides a single, trusted view of critical business data to users and applications. Reconciles overlapping, redundant, and inconsistent data from various systems.
- **Reference Data:** Provides a solution for centralized management, stewardship, and distribution of enterprise reference data. Supports defining and managing reference data as an enterprise standard.
- **Data Catalog:** Provides comprehensive capabilities to help understand information and foster collaboration between business and IT. Establishes a foundation for information integration and governance projects.
- **Data Models:** Comprehensive data models containing data warehouse design models, business terminology models and analysis templates to accelerate the development of business intelligence applications.
- **Data Quality Rules:** Establishes and manages high-quality data. Cleanses data and monitors data quality on an ongoing basis, helping to turn data into trusted information.

Security

The Security component is critical in all analytic and data architecture and designs. With a focus on data protection there are specific characteristics to keep in mind. These are the abilities to mask / hide data at a granular level for those that still need to interact with it, the ability to encrypt the data from all users from a static content perspective, the ability to know who accesses it and why, and the ability to have an overall view of all these activities.



- **Data Security** includes data encryption, data masking and data protection to protect all data in motion or at rest on all environments following the corporate security policy associated with PII data, to control user access to any data based upon their role and access level, and to be complaint with all industry and government regulations including SOX, HIPAA, ITAR, SOC2, and PCI.
- **Identity & Access Management** authenticates users, including IT administrators and application developers, for access to application and cloud resources. Users are assigned specific roles for access to different cloud resources or application instances.
- **Infrastructure Security** protects against network level threats and attacks with intrusion prevention and detection, including those

tunneling through encrypted web transactions and web applications deployed within the system. It protects virtual servers and applications against breaches. It tracks and supports regulatory compliance needs for the infrastructure, middleware, and workload.

- **Application Security** protects PaaS infrastructure against threats and attacks at the application level by using scans to determine if there are vulnerabilities before applications are deployed into production.
- **Secure DevOps** includes cloud information systems acquisition, development, and maintenance management of application and infrastructure resources. It includes vulnerabilities management and the associated mitigation of network and compute system vulnerabilities through patch management.
- **Security Monitoring & Intelligence** provides security and visibility into cloud infrastructures, data, and applications by collecting and analyzing logs in real time across the various components and services in the cloud. It provides real-time risk analysis of the workloads hosted in cloud against the myriad of known vulnerabilities and alerts against zero day attacks. It includes problem and information security incident management and responding to expected and unexpected events.
- **Security Governance** includes maintaining security policy and audit and compliance measures meeting corporate policies and market industry-specific regulations.

System Management

Cloud service management and operations refer to all of the activities that are performed by an organization to plan, design, deliver, operate, and control IT and cloud services that are offered to customers. The cloud provider Service Level Agreements (SLAs) may cover all details.

Deployment Considerations

The majority of the use cases for big data and analytics are hybrid cloud use cases. The CSCC has published a companion white paper, *Hybrid Cloud Deployment Considerations for Big Data and Analytics* that describes in detail the deployment considerations for a hybrid cloud strategy. [2]

Deployment Scenario

Now that we understand the architectural components of a big data analytics solution in the cloud, let's look at an example of how to use the solution components from this architecture. The use case below is about how a bank started their journey to become a cognitive bank.

Clients use a mobile application for all banking purposes. The bank uses this application for personalized interaction, including *next best action*. APIs are used within the branches for personalized client-centric interactions. The call center agents at the branches use a Q&A solution for credit cards using cognitive APIs. All interactions with the client are done through the interaction management system.

Cognitive insights are designed to help marketers understand and anticipate customer behaviors. It is important that the bank agents drive the correct communication with clients based on precise, real-time insights that can lead to treatments, financial tips, redirection, and more.

Business Challenge

Some of the significant challenges that are unique to the cognitive banking architecture implementation include:

- Ensure proper client insights by having a *segment of one* customer profile. This is the cornerstone of the digital strategy.
- Have an ecosystem of widgets to provide personalized interaction with the client, provide next best action and education, and harvest the communications with the client.
- Provide an interaction management system for all interactions using all of the channels with the client leveraging a cloud marketing platform.

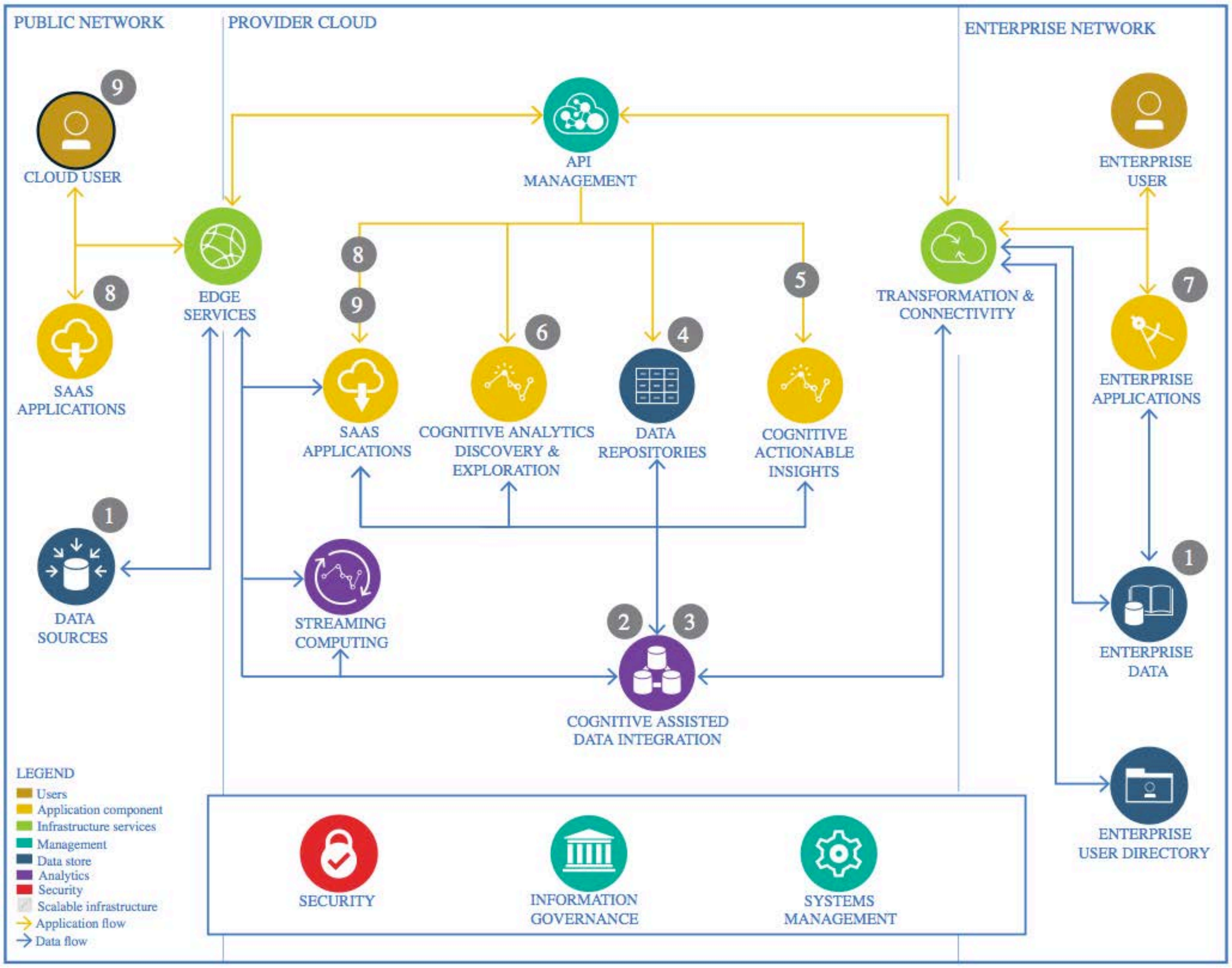


Figure 2: Solution Architecture for the Route to Cognitive Banking

Runtime Flow

1. Customer profile data (demographic, preferences, production services, historical behavior, and more) and interactions data from enterprise databases are collected in batch and real time. Public data such as social media data (Facebook & Twitter) is also collected.

2 & 3. Enterprise and public data are integrated and transformed using information services and Apache Spark.

4. Transformed data, which is a combination of structured customer data from enterprise data sources and interaction data that can vary in structure over a period of time, are stored on a JSON data repository in the data lake. Cached data for real-time API requests are stored on an in-memory database in the data lake, another data repository.

5. Perform rule-based analytics to analyze the personal client experience of a client who is at a branch or using the mobile application. Present changes to the branch personnel or mobile application where appropriate to improve the client experience in real time.

6. In the R&D environment, data scientists discover and explore patterns. Use cases underway include client insights, personal client experience, credit related query facilitation, private banking experience, and mobile application onboarding.

7. The APIs read the data upon real-time requests that are typically from branches.

8. A cloud marketing platform is used for all interactions with the client through all of the different channels, especially for the next best action or next best offer. For example, the mobile application takes in consideration social, localization, and mobile inputs to decide the best actionable insight in real time, using the power of information available to better engage with customers. This can lead to a better client experience and help move closer to *segment of one* customer profiling based on their individual context, behavior and preferences.

9. The cognitive APIs are used currently for Q&A for the credit card marketing agents.

References

[1] Designing and Operating a Data Reservoir, an IBM Redbook, Mandy Chessell, David Radley, Jay Limburn, Kevin Shank, Nigel L Jones.
http://www.redbooks.ibm.com/redbooks.nsf/RedpieceAbstracts/sg248274.html

[2] Cloud Standards Customer Council 2017, *Hybrid Cloud Considerations for Big Data and Analytics.*
http://www.cloud-council.org/deliverables/hybrid-cloud-considerations-for-big-data-and-analytics.htm

Appendix: Use Cases and Functional Requirements

Big data and analytics are transforming every industry. The table below shows common use cases and requirements for analytics in specific industries.

Industry	#	Use Case/Requirement	Detail
Telco	1	**Innovative business models**	Create data-driven API models for improved customer satisfaction.
	2	**Operational efficiency**	Offer world-class customer care by tracking in-depth subscriber activity, anticipating and monitoring issues and reducing call center wait times.
	3	**Real time analytics and decision making**	Control and reduce network congestion by profiling subscribers and dynamically allocating capacity.
Retail	4	**Personalized recommendations**	Create data-driven API models for improved personalized customer next best offers.
	5	**Dynamic pricing**	Provide differentiated dynamic pricing based on seasonal, fashion, and other trends.
	6	**In-store experience**	Use geo fencing based on near field communication (NFC) and provide guided selling experience in store.
Banking	7	**Fraud detection**	Enable real-time fraud detection, alerting, and remediation based on the transaction and behavioral data that is collected and analyzed for billions of transactions.
	8	**Sales and marketing campaigns**	Use comprehensive customer insights to provide targeted campaigns integrating them with personalized offers and schemes.
	9	**Threat detection and compliance**	Use big data analytics with active and passive probes to test system weaknesses and attacks, collect data for auditing, and provide a line-of-business dashboard of the overall exposure.

Healthcare and life sciences	10	**Summary of genomics**	Develop highly confident characterization of whole human genomes as reference materials to offer targeted care.
Government	11	**Census Bureau Statistical Survey response improvement**	Increase quality and reduce the cost of field surveys, even though survey responses are declining.

All of the use cases above deal with large volumes, velocities, and varieties of data. The following table addresses other functional capabilities expected from an analytics solution.

Requirement area	**Description**
Data sources	The analytics solution must support: • A wide variety of data sources and formats (such as csv, text, XML, JSON, images, and other formats), • A wide variety of data sizes (for example, data sets that may be very large) • The rate of growth • The processing of data at rest (stored) or in motion (in memory) The appropriate mechanisms for accessing data located in the data sources and delivering data to the data sources are dependent on the capabilities, capacity, and interfaces offered by these data sources. In practice, an analytics solution needs a range of data integration and provisioning capabilities to connect to the required data sources.
Data quality	The analytics solution must provide capabilities for cleansing, quality checking, pre-analytic processing, and more.
Data transformation	The analytics solution must be able to convert data from one format to another. The solution must also be able to correlate and match data from different sources for use by deployed analytics and other applications.
Capability infrastructure	The cloud provider should supply the tools to process data sets. These include platform tools that enable connectivity, load balancing, routing, and the like, or hardware resources such as suitable storage, compute, and networking.

Information Management	The cloud provider must provide multiple levels of data protection, data encryption, compliance rules, processes, and audit trails to meet privacy, sovereignty, and Intellectual Property (IP) protection guidelines for organizations in different industries.
Self-service discovery and exploration of data	The analytics solution must provide web-based, self-service analytics tools with capabilities like data exploration, discovery, ad-hoc Business Intelligence (BI) queries, and so. This empowers users to gain deeper insight from all kinds of data without imposing the need to explicitly define and maintain schemas.
Analytic model management	The analytics solution must offer a centralized repository for storing analytic models, so you can create, manage, validate, administer, and monitor analytic models.
Analytics deployment management and operation	The analytics solution must provide tools to develop, validate, combine multiple models, deploy, and retire analytic models, including the audit trail for model management, version control, and address model decay.
Metadata management	The analytics solution must provide end-to-end process, tools, and governance framework for creating, controlling, enhancing, attributing, defining, and managing a metadata schema, model, or other structured aggregation system.
Repository management	The analytics solution must support data modeling, data warehousing, data repositories, data integration, collections, and archiving.
Information visualization	The analytics solution must provide interactive graphical tools to explore and view data from all parts of the analytics solution.
Information governance	The analytics solution must support the policies, procedures, and controls that are implemented to manage information at an enterprise level in support of all relevant regulatory, legal, and risk requirements.

Acknowledgements

Major contributors to the whitepaper are: Christine Ouyang (IBM), Marcio Moura (IBM), Heather Kreger (IBM), Gopal Indurkhya (IBM), Anshu Kak (IBM), Mandy Chessell (IBM), Manav Gupta (IBM), Craig Statchuk (IBM) and Tracie Berardi (OMG).