Creating solutions and architectures that harness the value and power of big data and cloud can give your company a competitive advantage, spark new innovations, and increase revenues. The Cloud Standards Customer Council's Customer Cloud Architecture for Big Data and Analytics describes a well-tested and popular reference architecture for traditional analytics production environments. In this article, see how IBM supports this architecture in a secure, scalable, and flexible manner in dedicated and on-premises public cloud, hybrid cloud, and private cloud environments.

Introduction

Big data analytics 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.

The Cloud Standards Customer Council's *Customer Cloud Architecture for Big Data and Analytics* describes a well-tested and popular reference architecture for traditional analytics production environments. In this article, see how IBM supports the CSCC Customer Cloud Architecture for Big Data and Analytics in a secure, scalable, and flexible manner in dedicated and on-premises public clouds, hybrid cloud, and private cloud environments.

In addition:

- Discover business reasons for organizations to adopt cloud for their analytics needs.
- View 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 this solution.
- See how IBM products can be used to implement the architectural components.
- 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 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 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 advancement. 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.
- **Enhancement 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 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.

## Use cases and requirements

Big data and analytics are transforming every industry. Table 1 shows common use cases and requirements for analytics in specific industries.

### Table 1. Industry-specific use cases and requirements

<table>
<thead>
<tr>
<th>Industry</th>
<th>FR#</th>
<th>Use case/requirement</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telco</td>
<td>1</td>
<td>Innovative business models</td>
<td>Create data-driven API models for improved customer care.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Operational efficiency</td>
<td>Offer world-class customer care by tracking in-depth subscriber activity, anticipating and monitoring issues and reducing call center wait times.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Real time analytics and decision making</td>
<td>Control and reduce network congestion by profiling subscribers and dynamically allocating capacity.</td>
</tr>
<tr>
<td>Retail</td>
<td>4</td>
<td>Personalized recommendations</td>
<td>Create data-driven API models for improved customer care.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Dynamic pricing</td>
<td>Provide differentiated dynamic pricing based on seasonal, fashion, and other trends.</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>In-store experience</td>
<td>Use geo fencing based on near field communication (NFC) and provide guided selling experience in store.</td>
</tr>
<tr>
<td>Banking</td>
<td>7</td>
<td>Fraud detection</td>
<td>Enable real-time fraud detection, alerting, and remediation based on the transaction and behavioral data that is collected and analyzed for billions of transactions.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Sales and marketing campaigns</td>
<td>Use comprehensive customer insights to provide targeted campaigns integrating them with curated offers and schemes.</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Threat detection and compliance</td>
<td>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.</td>
</tr>
<tr>
<td>Healthcare and life sciences</td>
<td>10</td>
<td>Summary of genomics</td>
<td>Develop highly confident characterization of whole human genomes as reference materials to offer targeted care.</td>
</tr>
<tr>
<td>Government</td>
<td>11</td>
<td>Census Bureau Statistical Survey response improvement</td>
<td>Increase quality and reduce the cost of field surveys, even though survey responses are declining.</td>
</tr>
</tbody>
</table>
All of the use cases in Table 1 deal with large volumes, velocities, and varieties of data. Table 2 addresses other functional capabilities expected from an analytics solution.

**Table 2. Additional function capability requirements for an analytics solution**

<table>
<thead>
<tr>
<th>Requirement area</th>
<th>Description</th>
</tr>
</thead>
</table>
| Data sources                              | The analytics solution must support:  
• A wide variety of data sources and formats (such as CSV, text, XML, 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, transformation              | The analytics solution must provide capabilities for cleansing, converting, 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.          |
| Security and data protection              | The cloud provider must provide multiple levels of security, compliance rules, processes, and audit trails to meet privacy, sovereignty, and 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 BI queries, and so on. 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. |
Non-functional requirements

Non-functional requirements dictate the design of a big data analytics solution since data must be provisioned and structured on multiple data platforms to support the different analytics workloads. These non-functional requirements are in addition to the typical non-functional requirements that any cloud-based solution needs to be viable. Table 3 summarizes the key categories for most non-functional requirements.

Table 3. Key categories for non-functional requirements

<table>
<thead>
<tr>
<th>Non-functional requirements</th>
<th>Typical considerations</th>
</tr>
</thead>
</table>
| Performance                        | • Speed, latency, throughput of requests and data transfers to and from the cloud provider  
• Matching of data location and structure to workload requirements |
| Security                           | • User management, access, authentication, and authorization, as well as integration with security systems within enterprise network  
• Integrity of data and services offered by the cloud provider  
• Data authorization under the control of the data owner and driven by rules and metadata  
• Auditability of all actions performed by users and cloud administrators  
• Audit data available for analytics |
| Operational and environmental      | • Release cycles (consider the expected frequency and impact of future releases of the product) of the cloud provider capabilities  
• Resilience and disaster recovery offered by cloud provider  
• Integration points with enterprise systems since data must freely flow both into and out of the analytics solution; the analytics solution must preserve the rights and access needs of the data owner. |
| Maintainability                    | • Adaptability and ongoing development (consider likely future changes)  
• Well-defined outage windows enabling enterprises to take action and minimize impact |
| User experience                    | • Capability to customize appearance and branding of offered capabilities  
• Internationalization and other foreign market considerations |
| Legal / Compliance                 | • Compliance with regulations, industry standards  
• Certification from third parties  
• Use of licensed components and data sets |
| Resilience and capacity management | • Reliability, availability, and fault-tolerance of the underlying infrastructure and capabilities offered by the cloud provider  
• Capacity (for example, volumes of data to be held or numbers of concurrent users) on offer |
| Scalability                        | • The ability of the system to adapt based on increasing or decreasing demands |

Architecture

Use the big data and analytics and cloud architecture guidance in this article to understand proven architecture patterns that have been deployed in numerous successful enterprise projects by IBM. Learn how to implement the architectures using IBM products and business partners.
Cloud deployments offer a choice of private, public, and hybrid architectures. Private cloud employs in-house data and processing components running behind corporate firewalls. 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 and public 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, that provides 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, that provides 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, that gives you the responsibility for connecting data sources to the cloud solution. The cloud 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.

Being able to choose a cloud deployment is important because selecting the data and processing location is one of the first architectural decisions for an analytics cloud project. The ability to select the cloud deployment and location allows both flexibility in operating models and the optimal placement of both the 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.

Choosing a cloud architecture allows compute components to be moved near data to optimize processing when data volume and bandwidth limitations produce remote data bottlenecks.

Figure 1 shows simplified enterprise cloud architecture for a big data and analytics production environment. The architecture has four three-network zones: public networks, provider clouds, and enterprise networks.
This big data architecture in a cloud-computing environment has many similarities to a traditional data warehouse deployment in a data center. Data is collected from structured and non-structured data sources. Data integration and stream computing engines stage and transform the data through different data repositories. The data is transformed, augmented with analytical insight, correlated, and summarized as it is copied and moved through this processing chain. Selected data is made available to consumers through APIs.

In each processing phase, information governance and security subsystems define and enable regulation and policies for all data across the system. Compliance is tracked to ensure controls are delivering expected results. Security covers all elements including generated data and analytics.

Users

Users of the analytics cloud solution can perform different roles, including:

- Data analysts who perform tasks related to collecting, organizing, and interpreting information. In a cloud computing environment, such users typically access information from streaming or data repositories, and make decisions on mechanics of data integration (such as the type of data integration services that should be used, the type of cleansing that needs to be performed, and other similar decisions).
- Data scientists who extract knowledge from data by leveraging their strong foundation in computer science, data modeling, statistics, analytics, and math. Data scientists are part analyst, part artist. They sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or solve a pressing business problem.
- Business users who are interested in information that will enable them to make decisions that are critical to tactical and strategic business operations.
- Solution architects who are responsible for identifying the components needed from the cloud provider in order to solve business problems.
These users 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 assist by preparing 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 further restricted as dictated by corporate policy and controlled by the appropriate data owners.

**Data lake architectures**

Many times, business units within an organization need access to similar data, although they may need to structure it differently to support their specific processing. As the volume of data grows, there is a desire to share the same copy of data to avoid both storage and network costs. The need for speed and agility within the business units creates a demand for immediate, self-service access to a wide variety of data from both inside and outside of the business. This immediate, self-service access enables the users to quickly make decisions based on facts and analytical insight and to rapidly develop new business functions as a result.

To address these requirements, the traditional way of deploying analytics has evolved to a data lake-based architecture that serves as a DevOps environment for data, collaboration, and analytics. The data lake adds common metadata and semantic definitions to descriptions of enterprise data repositories that are stored in a catalog. These catalog entries are augmented with governance classifications, rules, and policies that the processing engines use 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.

The catalog within the data lake in the cloud enhances the operation of the analytical solution with the following capabilities:

- Provides consumers access to the data they need, on demand through self-service interfaces that locate and deliver data to a wide range of analysis tools.
- Collaboration around data, analytics models, and the resulting insight.
- Support for data owners to manage and govern their data without IT intervention.

The data lake in the cloud enables the full analytics lifecycle by allowing:

- **Search, discovery, and survey**: Access descriptions of data and analytics resources in a catalog to find and gain an understanding of the data and related assets that are available.
- **Analytics exploration**: Provision sandboxes of interesting data, prepare it for analysis, and create analytical models.
- **Analytical deployment**: Deploy analytical models to the analytic engines, data repositories, data integration, and stream computing engines to automate their execution as part of the organization's operations.
Information governance and security are two very important items to consider when deploying a data lake in the cloud.

The information governance for a data lake in the cloud is based on metadata management, information curation, classification schemes, policy and standards, rules, processes, exception management, reporting, and auditing.

The information security for a data lake in the cloud is based on:

- Data curation for security and protection
- Data access approval by a subject-area owner
- Well-defined access points
- Data-centric security access
- Isolated repositories
- Access monitoring and logging
- Security analytics and investigation
- Security audit and review

The following figure gives you a high-level overview for a data lake in the cloud.

**Figure 2. Data lake in the cloud: A high-level view**

The strong symmetry in the data-lake architecture reflects that the data lake is a hub solution, with data entering and leaving from both the enterprise and from the public network. The ad hoc nature of many of the user interactions with the data lake and the increased variety of data that it stores means that the flow of data through to the repositories is tightly controlled by the APIs and data integration components. These components are closely integrated with information governance, which, in turn, makes use of the catalog to determine which governance actions to take.

**Component model**

The following sections describe each of the major components, the capabilities for a data lake in the cloud, and how IBM supports them. For a detailed explanation of the components in the
original big data and analytics architecture (shown in Figure 1), see CSCC Customer Cloud Architecture for Big Data and Analytics.

**Public network components**

The public network contains elements that exist in the Internet: data sources and APIs, users, and the edge services needed to access the provider cloud or enterprise network.

**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.

**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. Such applications include:

- **Customer experience**: Customer-facing cloud systems can be a primary system of engagement that drives new business and offers existing clients 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**: As data is consolidated and reported faster and easier than ever before, the office of finance should become more efficient.
- **Risk**: Having more data available across a wider domain means that risk analytics are more effective. Elastic resource management increases the processing power in times of heightened threat.
- **IT economics**: IT operations are streamlined as capital expenditures are reduced while performance and features are improved by cloud deployments.
- **Operations and fraud**: Cloud solutions can provide faster access to more data, which allows for more accurate analytics that flag suspicious activity and offer remediation in a timely manner.

The data within these applications can be valuable to the data lake. These applications may also receive insight from the data lake or be used as the deployment platform for real-time analytics models developed in the data lake.

**Public sources**

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

In a typical big data system, there can be a number of different information sources, 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. Big data tools enable organizations to use this data, but these tools are typically run on-premises and often require substantial up-front investment. Cloud computing helps mitigate that investment and the associated risk by providing big data tools via a pay-per-use model. Data sources include:

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

IBM supports any of these data sources as input into the analytics process as well as data from enterprise data stores, cloud data stores, and enterprise applications.

**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 comes from the external Internet, the flow can 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 Servers**: 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.
- **Firewalls**: 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.

IBM® Bluemix™ offers the Secure Gateway service to connect Bluemix applications to remote locations on-premises or in the cloud. The Secure Gateway provides secure connectivity and establishes a tunnel between a client's Bluemix organization (cloud provider network) and the remote location (enterprise network) that you want to connect to.

The IBM Virtual Private Network (VPN) service provides a secure IP-layer connectivity between your on-premise data center and your Bluemix cloud. It leverages Internet Protocol Security (IPsec) protocol suite for protecting IP communication between endpoints residing on your private subnets. An IPsec-compatible VPN gateway is required in your on-premise data center for establishing secure connectivity with IBM VPN service.

IBM SoftLayer provides secure connectivity to the SoftLayer private network over SSL, PPTP, or IPSEC VPN gateways. SoftLayer also provides on-demand Content Delivery Networks (CDN) to distribute content geographically. SoftLayer load-balancing solutions are configurable and flexible so you can manage the traffic and resource usage of server nodes in your environment. Local, global, and high-availability options can be activated, changed, and deactivated at any time.

**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
- Data integration
- Data repositories
- Analytics discovery and exploration
- Deployed analytics
- Security
- Information governance
- Transformation and connectivity

**API management**

API management components manage a catalog of APIs from a wide variety of deployment environments. The catalog enables developers and end users to locate the APIs they need to rapidly assemble solutions.

The IBM API Management family of products gives you software tools to create, manage, and share APIs in a secure, scalable environment.
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:

- **Real-time analytical processing**: Applying analytic processing and decision making to in-memory and transient data with minimal latency.
- **Data augmentation**: Filtering and diverting in-motion data to warehouses for deeper background analysis.

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

IBM Streams is an advanced data stream-processing platform that allows user-developed applications to quickly ingest, analyze, and correlate information as it arrives from thousands of data stream sources. The solution can handle high data throughput rates—up to millions of events or messages per second.

Real-time analysis on data in motion can be performed on Bluemix by using the IBM Streaming Analytics service.

Data integration

Data integration copies and correlates information from disparate sources to produce meaningful associations related to primary business dimensions. A complete data integration solution encompasses access connectors, discovery of data source characteristics, cleansing, monitoring, transforming, and delivering data. Information provisioning methods include ETL (extract, transform, and load), event-based processing, services, federation, change data capture with replication, and continuous-stream ingestion.

Data to be integrated can come from public network data sources, enterprise data sources, or streaming computing results. The results from data integration can feed streaming computing, 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:

- **Data staging**: Converting data to the appropriate formats for downstream processing.
• **Data quality**: Cleaning and organizing data to remove redundancies and inconsistencies so that it meets the quality needs of the data consumers.

• **Provisioning**: Transforming, performing governance actions, and delivering data to the appropriate destinations. Provisioning may move data between data repositories or to and from the data sources.

• **Entity services**: Matches data together from different sources to build a more complete view of key entities such as customers, products, and assets.

The IBM products for data integration include:

• **IBM BigInsights® for Apache Hadoop** is a mature, Hadoop-based solution for big data analytics.

• **IBM InfoSphere® Information Server** is a data integration and governance platform that helps you understand, cleanse, transform, and deliver trusted information to your critical business initiatives, such as big data, master data management, and point-of-impact analytics.

• **IBM InfoSphere Data Replication** is a data replication platform for maintaining copies of data in different data stores. It is easy to use, highly scalable, and enterprise ready. It can provide trusted data synchronization (including change data capture capabilities) to replicate information between heterogeneous data stores in near real time.

• **IBM InfoSphere Big Match for Hadoop** helps you match values from big volumes of structured and unstructured data to derive deeper insight into the characteristics of particular entities. The scalable platform can enable fast, efficient linking of data sources to provide complete and accurate customer information—without increasing risk of errors or data loss when moving data from source to source.

**Data repositories**

The data stored in the cloud environment is organized into repositories. These repositories may be hosted on different data platforms (such as a database server, Hadoop, or a NoSQL data platform) that are tuned to support the types of analytics workload that is accessing the data.

The data that is stored in the repositories may come from legacy, new, and streaming sources, enterprise applications, enterprise data, cleansed and reference data, as well as output from streaming analytics.

Types of data repositories include:

• **Catalog**: Results from discovery and IT data curation create a consolidated view of information that is reflected in a catalog. The introduction of big data increases the need for catalogs that describe what data is stored, its classification, ownership, and related information governance definitions. From this catalog, you can control the usage of the data.

• **Data virtualization**: Agile approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data.
• **Landing, exploration, and archive**: Allows for large datasets to be stored, explored, and augmented using a wide variety of tools since massive and unstructured datasets may mean that it is no longer feasible to design the data set before entering any data. Data may be used for archival purposes with improved availability and resiliency thanks to multiple copies distributed across commodity storage.

• **Deep analytics and modeling**: The application of statistical models to yield information from large data sets comprised of both unstructured and semi-structured elements. Deep analysis involves precisely targeted and complex queries with results measured in petabytes and exabytes. Requirements for real-time or near-real-time responses are becoming more common.

• **Interactive analysis and reporting**: Tools to answer business and operations questions over Internet-scale data sets. Tools also use popular spreadsheet interfaces for self-service data access and visualization. APIs implemented by data repositories allow output to be efficiently consumed by applications.

• **Data warehousing**: Populates relational databases that are designed for building a correlated view of business operation. A data warehouse usually contains historical and summary data derived from transaction data but can also integrate data from other sources. Warehouses typically store subject-oriented, non-volatile, time-series data used for corporate decision-making. Workloads are query intensive, accessing millions of records to facilitate scans, joins, and aggregations. Query throughput and response times are generally a priority.

IBM offers a wide variety of offerings for consideration in building data repositories:

• **InfoSphere Information Governance Catalog** maintains a repository to support the catalog of the data lake. This repository can be accessed through APIs and can be used to understand and analyze the types of data stored in the other data repositories.

• **IBM InfoSphere Federation Server** creates consolidated information views of your data to support key business processes and decisions.

• **IBM BigInsights for Apache Hadoop** delivers key capabilities to accelerate the time to value for a data science team, which includes business analysts, data architects, and data scientists.

• **IBM PureData™ System for Analytics**, powered by Netezza technology, is changing the game for data warehouse appliances by unlocking data’s true potential. The new IBM PureData System for Analytics is an integral part of a logical data warehouse.

• **IBM Analytics for Apache Spark** is a fully-managed Spark service that can help simplify advanced analytics and speed development.

• **IBM BLU Acceleration®** is a revolutionary, simple-to-use, in-memory technology that is designed for high-performance analytics and data-intensive reporting.

• **IBM PureData System for Operational Analytics** is an expert integrated data system optimized specifically for the demands of an operational analytics workload. A complete solution for operational analytics, the system provides both the simplicity of an appliance and the flexibility of a custom solution.

Bluemix offers a wide variety of services for data repositories:
• **BigInsights for Apache Hadoop** provisions enterprise-scale, multi-node big data clusters on the IBM SoftLayer cloud. Once provisioned, these clusters can be managed and accessed from this same service.

• **Cloudant® NoSQL Database** is a NoSQL Database as a Service (DBaaS). It’s built from the ground up to scale globally, run non-stop, and handle a wide variety of data types like JSON, full-text, and geospatial. Cloudant NoSQL DB is an operational data store optimized to handle concurrent reads and writes and provide high availability and data durability.

• **dashDB™** stores relational data, including special types such as geospatial data. Then analyze that data with SQL or advanced built-in analytics like predictive analytics and data mining, analytics with R, and geospatial analytics. You can leverage the in-memory database technology to use both columnar and row-based tables. The dashDB web console handles common data management tasks, such as loading data, and analytics tasks like running queries and R scripts.

**Analytics discovery and exploration**

Analytics discovery and exploration supports the development of new analytics models. It includes the following types of services:

• **Self-service**: Enables users to sign up, access the output from analytics systems, and customize analytical processing. The user may be an employee of the enterprise, the cloud provider, or some other third party.

• **Visualization**: Enables users to create and use dashboards to explore and interact with data from the data repositories, actionable insight applications, or enterprise applications. The user must be authorized to access the visualization.

• **Data preparation**: Enables users to transform data from its raw form to a format that is easier to analyze.

• **Sandbox**: Provides the ability to copy samples of data sets into a private area for experimentation.

**Watson® Analytics** is an IBM offering for the components listed in analytics discovery and exploration. Watson Analytics offers the benefits of advanced analytics without the complexity. A smart data discovery service available on the cloud, it guides data exploration, automates predictive analytics, and enables effortless dashboard and infographic creation.

**IBM DataWorks** is a Bluemix service for analytics discovery and exploration. DataWorks is a fully managed data preparation and movement service available on Bluemix that enables business analysts, developers, data scientists, and engineers to put data to work. DataWorks empowers technical and non-technical users to discover, cleanse, standardize, transform, and move data in support of application development and analytic use cases.
Deployed analytics

Deployed analytics is the collection of analytics applications that are deployed on data repositories to extract value and insights to ultimately derive actions from data. There are a number of deployed analytics applications available today, including:

- Decision management includes analytics-based decision management that enables organizations to make automated decisions backed by analytics, improve efficiency and enable collaboration. Decision management applications also include operational decision-management systems that rely on rules to augment enterprise decision making to achieve specific business objectives (such as prevent a customer from churning, converting a visitor to a client, ordering more inventory, and more).
- Predictive analytics services extract information from existing data sets to determine the current state, identify patterns, and predict future trends.
- Analysis and reporting about operational and warehouse data to business stakeholders and regulators where big data typically increases the scope and depth of available data.
- Content analytics services enable businesses to gain insight and understanding from their structured and unstructured content (also referred to as textual data). A large percentage of the information in a company is maintained as unstructured content, such as documents, blobs of text in database, wikis, and other content.
- Planning and forecasting enables faster and more efficient development of plans, budgets, and forecasts by creating, comparing, and evaluating business scenarios.
- Visualizations are often needed to share an idea or drive consensus in a way that simplifies the complex associations combined with high data volumes.

The IBM offerings that provide deployed analytics services include:

- **IBM Operational Decision Management** is a full-featured, easy-to-use platform for capturing, automating, and governing frequent, repeatable business decisions. It consists of two components, IBM Decision Center and IBM Decision Server. They form the platform for managing and executing business rules and business events to help you make decisions faster, improve responsiveness, minimize risks, and seize opportunities.
- **IBM SPSS®** predictive analytics software enables you to predict with confidence what will happen next so that you can make smarter decisions, solve problems, and improve outcomes.
- **IBM Cognos®** software enables organizations to become top-performing and analytics-driven entities. From business intelligence to financial performance and strategy management, Cognos software is designed to help everyone in your organization make the decisions that achieve better business outcomes—today and in the future.
- **IBM Content Analytics** leverages the sophisticated natural language processing (NLP) technology that helps organizations aggregate, analyze, and visualize massive amounts of information to expose unique insights. It helps organizations interpret and understand their enterprise information to validate what is known and reveal what is unknown.
Bluemix services that offer deployed analytics capabilities include:

- **The Business Rules** service enables developers to spend less time recoding and testing when the business policy changes. The Business Rules service minimizes your code changes by keeping business logic separate from application logic.
- **Watson services** is a collection of services that provide capabilities such as concept expansion, entity analytics, language translation, document conversion, and similar capabilities.
- The **Predictive Analytics** service is a full-service Bluemix offering that makes it easy for developers and data scientists to work together to integrate predictive capabilities with their applications. Built on IBM's proven SPSS analytics platform, Predictive Analytics allows you to develop applications that make smarter decisions, solve tough problems, and improve user outcomes.
- The **Embeddable Reporting** service provides dashboards and reports for web or mobile applications. It provides a mechanism to author reports using a simple cloud editor, enabling users to embed reports and dashboards using a variety of languages such as Node.js or Java using a RESTful API.

**Security**

Rigorous security is needed at each step in the lifecycle of big data—from raw input sources to valuable insights to sharing of data among many users and application components. Security services enable identity and access management, protection of data and applications, and actionable security intelligence across cloud and enterprise environments. It uses the catalog to understand the location and classification of the data it is protecting.

Identity and access management components enable authentication and authorization (access management), as well as privileged identity management. Access management ensures each user is authenticated and has the right access to the environment to perform their task based on their role (that is, data analysts, data scientists, business users, solution architects). Capabilities should include granular access control (giving users more precision for sharing data) and single sign-on facility across big data sources and repositories, data integration, data transformation, and analytics components.

Privileged identity management capabilities protect, automate, and audit the use of privileged identities to ensure that the access rights are being used by the proper roles, to thwart insider threats, and to improve security across the extended enterprise, including cloud environments. This capability generally uses an enterprise user directory.

Application and data protection services enable and support data encryption, infrastructure and network protection, application security, data activity monitoring, and data lineage where:

- Data encryption secures the data interchange between components to achieve confidentiality and integrity with robust encryption of data at rest as well as data in transit.
• Infrastructure and network protection supports the ability to monitor the traffic and communication between the different nodes (like distributed analytical processing nodes) as well as prevent man-in-the-middle, disk operating system attacks. This will also send alerts about the presence of any bad actors or nodes in the environment.

• Application security ensures security is part of the development, delivery, and execution of application components, including tools to secure and scan applications as part of the application development lifecycle. Application security identifies and remedies security vulnerabilities from components that access critical data before they are deployed into production.

• Data activity monitoring tracks all queries submitted and maintains an audit trail for all queries run by a job. The component provides reports on sensitive data access to understand who is accessing which objects in the data sources.

• Data lineage traces the origin, ownership, and accuracy of the data and complements audit logs for compliance requirements.

Security intelligence enables security-information event management, audit and compliance support for comprehensive visibility, and actionable intelligence to detect and defend against threats through analysis of events and logs. High-risk threats that are detected can be integrated with enterprise incident management processes. This component enables the audit capability to show that the analytics delivered by the big data platform sufficiently protects PII and delivers anonymity as well as enabling automated regulatory compliance reporting.

IBM capabilities for the security components include:

• **IBM Security identity and access management solutions** strengthen compliance and reduce risk by protecting and monitoring user access in today's multi-perimeter environments. These solutions safeguard mobile, cloud, and social access, prevent advanced insider threats, simplify cloud integrations and identity silos, and deliver actionable identity intelligence.

• **IBM Security Guardium** is a comprehensive data security platform that provides a full range of capabilities to protect sensitive data— from discovery and classification of sensitive data to vulnerability assessment to data and file activity monitoring to masking, encryption, blocking, alerting, and quarantining.

• **IBM security intelligence and analytics solutions** collect security-relevant information and apply advanced analytics and automation to protect against threats.

Bluemix offers the **Application Security Manager** security intelligence service that provides a set of capabilities to enable organizations to take a strategic, risk-based approach to the application security problem.

Additionally, you can implement user authentication for your web and mobile apps quickly, using simple policy-based configurations.

**Information governance**

Information governance provides the policies and capabilities that enable the analytics environment to move, manage, and govern data. Information governance provides:
• Management interfaces to enable the business team to control and operate the processes that manage data
• Protection classification and rules for managing and monitoring access, masking, and encryption
• Workflows for coordinating changes to the data repositories, catalog, data, and supporting infrastructure between different teams

IBM capabilities for information governance are as follows:

• **InfoSphere Information Governance Catalog** provides comprehensive capabilities to define the information governance program's policies, classifications, and rules, catalog the data sets available to the organization, and use analytics to help understand and govern information flow within an organization. By defining a common business language, InfoSphere Information Governance Catalog encourages a standardized approach to managing data and aligns business and IT data requirements.

• **IBM Security Guardium** is a comprehensive data security platform that provides a full range of capabilities – from discovery and classification of sensitive data to vulnerability assessment to data and file activity monitoring to masking, encryption, blocking, alerting, and quarantining sensitive data.

• InfoSphere Optim Data Management solutions manage data protection from requirements to retirement.

**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 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.

IBM capabilities for transformation and connectivity include:

• **IBM DataPower® Gateway** is a multichannel gateway platform to secure, integrate, control, and optimize delivery of workloads across multiple channels including: mobile, API, web, SOA, business to business (B2B), and cloud.

• **IBM Integration Bus** is a market-leading enterprise service bus (ESB) that offers a fast, simple way for systems and applications to communicate with each other. It supports a range
of integration choices, skills, and interfaces to optimize the value of existing technology investments. It provides the ability to perform business transaction monitoring (BTM) and is a vital platform for the API economy and analytics.

- IBM Aspera® offers transfer software that moves the world's data at maximum speed, regardless of file size, transfer distance, or network conditions.

Bluemix has the following services in this area:

- The Secure Gateway service brings hybrid integration capability to your Bluemix environment. It provides secure connectivity from Bluemix to other applications and data sources running on-premises or in other clouds. A remote client is provided to enable secure connectivity.
- The API Management service enables developers and organizations to manage and enforce policies around the consumption of their business services. Use an existing API, or design a new API. Then apply security controls, set rate limits, test APIs in place, and finally publish these "managed APIs" on Bluemix--either to you, to select developer organizations, or to app developers outside of Bluemix. Share your APIs using an available self-service portal that can be white-labeled and provides built-in support for blogs, discussion forums, comments, ratings, FAQs, and the APIs that you choose to publish. This service includes API versioning, lifecycle management, and API usage analytics.

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

Enterprise users
Enterprise users are individuals that connect to the analytics cloud solution through the organization’s internal network. Users set up or use the results of the analytical system, and are typically part of the enterprise. Users can be:

- Administrative users, setting up the analytical processing system.
- Analytical services users, using the results of the analytical system.
- Enterprise users, invoking enterprise applications in the analytical system. In the case of enterprise users, the access path might not go through the public Internet and may go directly to the analytical insights or enterprise applications.

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

Applications include:

- **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**: As data is consolidated and reported faster and easier than before, the office of finance should become more efficient.

• **Risk**: Having more data available across a wider domain makes risk analytics more effective. Elastic resource management means more processing power is available in times of heightened threat.

• **IT economics**: Cloud deployments streamline: IT operations, which reduces capital expenditures, while improving performance and features.

• **Operations and fraud**: Cloud solutions can provide faster access to more data which allows for more accurate analytics that flag suspicious activity and offer remediation in a timely manner.

IBM provides a number of enterprise applications either as SaaS offerings or as traditional software.

**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 enterprise applications running in the enterprise. It is a blend 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 document 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 have been generated from, and other contextual information.

• **Historical data**: Data from past analytics and enterprise applications and systems. Historical data may be an archive of a system or data from a system that has been decommissioned.

IBM capabilities for hosting enterprise data are the same as those for data repositories described in Provider cloud components.

**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.

**Complete picture**

The following figure provides a more detailed architectural view of components, subcomponents, and relationships for a cloud-based analytics solution that provides traditional historical analysis of an organization’s data.

**Figure 3. Architectural overview for big data analytics using cloud with subcomponents**

The figure below provides a detailed architectural view of components, subcomponents, and relationships for data lakes running in cloud environments.
Figure 4. Architectural overview for data lakes using cloud with subcomponents

IBM product support for big data and analytics solutions in the cloud

Now that we've reviewed the component model for a big data and analytics solution in the cloud, let's look at how IBM products can be used to implement a big data and analytics solution. In previous sections, we highlighted IBM's end-to-end solution for deploying a big data and analytics solution in cloud.

The figure below shows how IBM products map to specific components in the reference architecture.
Figure 5. IBM product mapping

Bluemix services support for the capabilities
The figure below shows how IBM products map to specific components in the reference architecture.

Figure 6. Bluemix services mapping

IBM product support for data lakes using cloud architecture capabilities
The following images show how IBM products can be used to implement a data lake solution. In previous sections, we highlighted IBM's end-to-end solution for deploying data lake solutions using cloud computing.

Mapping on-premises and SoftLayer products to specific capabilities
Figure 7 shows how IBM products can be used to run a data lake in the cloud.
Scenarios

Now that you understand the architectural components of a big data analytics solution in the cloud, let's look at how to use IBM products to implement common scenarios using this architecture. We'll showcase actual business scenarios and cover an example deployment configuration for that scenario.

- **Scenario 1. Fraud and identity theft**
- **Scenario 2. Cyber threat intelligence solution for telecommunication companies (telcos)**
- **Scenario 3. Innovative business models for telcos**
• **Scenario 4. Predictive customer intelligence and behavior-based customer insights**

These scenarios reuse the components that the organization is currently using in their traditional data centers, which we depict as part of the enterprise zone of the architecture.

**Scenario 1. Fraud and identity theft analytics applications**

Figure 9 shows the flow of a typical use case for fraud and identity theft analytics applications.

In this example, a compliance and security analyst is investigating fraud and identity theft threats related to banking operations. The yellow flows show the interactions of the compliance officers, while the blue flows show the flow of data across the analytical system.

**Figure 9. Fraud and identity theft basic information flow**

The steps in this process are as follows:

1. To detect identity theft and correlate financial activity, enterprise compliance officers customize and configure the analytical processing system on the cloud provider to look at banking transaction data from the enterprise as well as social media feeds from the public network.
2. Data flows from public data sources (like social media) through edge services which route the data to the data integration components in the provider cloud.
3. Data integration from components such as IBM Information Server is used to extract data from bank transactions, credit applications, user name, and address changes along with financial information from related institutions. Social media feeds are harvested for current location and activities. Collected and correlated data is enriched with directory information stored on-premises to associate bank account information to past, current, and new customers. Enterprise data stores are augmented with summary data as required by dependent applications.
4. Credit card transactions are forwarded directly to streaming computing such as InfoSphere Streams. In some cases, correlation of streaming data with other information is used to flag outliers and other potential threats. For example, user names need to be enriched with the
last known location (perhaps from social media) to provide alerts about the same customer being in more than one place at one time.

5. Incoming data from structured and streaming sources, along with related streaming analytics, are cached in the landing, exploration, and archive component within data repositories such as IBM BigInsights for Apache Hadoop. Other data is largely historical in nature. It requires complex, multi-pass machine learning algorithms to detect and flag unusual behavior. One example is entity analytics which seeks to distinguish clients with the same name and alternatively highlight people with different web identifiers, like email addresses and user names that actually represent the same individual.

6. Data that is flagged for further investigation is investigated by a case management team that run ad hoc analytics against new and historical data to find outliers and other abnormal behavior. The result of this analysis is fed back into the process and enterprise applications to capture subsequent instances of fraud.

7. After data has been collected, cleansed, transformed, and stored, it is communicated to a decision-management application such as IBM Operational Decision Manager which determines whether a case should be opened for further investigation and action by the Fraud and Identity Theft team using a case management tool such as IBM Case Manager. Many other types of analytics may be deployed, such as predictive analytics to classify incoming transactions against an established profile and flag potential outliers that represent identity theft threats.

8. At the end of the analytical process, enterprise users, such as compliance officers, use visualizations and interactive tools to provide alternative views of data and analytics. They promote better understanding of results by showing important areas of interest, highlighting outliers, offering innovative ways to refine and filter complex data, and by encouraging deeper exploration and discovery. Some applications and related data may be made available to third-party users who access the enterprise applications via edge services which, in turn, collaborate with security services and the enterprise user directory.

Cloud architecture makes this type of solution easier to implement and maintain. As demand increases, more resources must be acquired. The introduction of feedback loops to introduce new analytics is made easier by cloud APIs that formalize the interactions between components. The continuous flow of data and updating of applications means that users can get the latest upgrades faster and easier.

### Hybrid cloud deployment example

The following table shows an example deployment of the above architecture in a hybrid cloud deployment model.

<table>
<thead>
<tr>
<th>Architectural component</th>
<th>Capability used</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services component</td>
<td>Enterprise edge services</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Streaming computing</td>
<td>InfoSphere Streams</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Data integration</td>
<td>InfoSphere Information Server</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Actionable insight</td>
<td>IBM Case Manager</td>
<td>SoftLayer</td>
</tr>
</tbody>
</table>

**Table 4. Fraud and identify theft analytics in a hybrid cloud deployment model**
Scenario 2. Cyber threat intelligence solution for telecommunication companies

A telecom provider needs a way to combat cyber threats in real time and stop advanced persistent threats that take longer to manifest. The following figure shows a typical cyber threat intelligence solution for a telecommunications company.

Figure 10. Cyber threat intelligence solution

The steps in a typical cyber threat intelligence solution are as follows:

1. Data is collected by both internal sources (such as network probes, DNS, NetFlow, AD logs, and network logs) and by external sources (such as blacklist and whitelist providers).
2. Most structured sources of data are sent first to the Security Information and Event Manager (SIEM) which acts as a data integration layer and converts all incoming data into a single format.
3. InfoSphere Streams (stream computing) picks up both the streaming flow data (such as DNS and NetFlow) as well as processed data from the SIEM system. It then computes simple analytics (such as traffic in/out per server, number of requests made/failed to a DNS domain, etc) which are used in developing machine-learning models.
4. All raw data and output from IBM Streams is sent to the data repository (stored on IBM BigInsights for Apache Hadoop in this case). Machine-learning models are run against longer data sets to detect advanced persistent threats. Additional models from stats language R are also deployed.
5. Machine-learning models that have been developed in SPSS are deployed in InfoSphere Streams, which scores them in real time to analyze network, user, and traffic behavior.
6. Custom blacklists from the client and other data sources such as AD logs are used to enrich and pinpoint user activity. Security analysts use i2 Analyst notebook BigSheets interface supplied by BigInsights for Apache Hadoop for visualizations.
7. User look-up information is ingested from enterprise Active Directory to establish exactly which user was involved in a particular traffic flow.

**Hybrid cloud deployment example**

Table 5 shows an example deployment of the above architecture in a hybrid cloud deployment model.

**Table 5. Cyber threat intelligence solution in a hybrid cloud deployment model**

<table>
<thead>
<tr>
<th>Architectural component</th>
<th>Capability used</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services component</td>
<td>Enterprise</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Streaming computing</td>
<td>InfoSphere Streams</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Data integration</td>
<td>Client's existing SIEM</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Data repositories</td>
<td>IBM BigInsights for Apache Hadoop</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Actionable insight</td>
<td>IBM SPSS</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Enterprise data</td>
<td>InfoSphere Master Data Management</td>
<td>Traditional data center</td>
</tr>
</tbody>
</table>

**Scenario 3. Innovative business models for telcos**

A telecommunication company is collecting call detail records (CDR) for billing purposes. This client wants to monetize the data it's collecting by combining existing data with other sources, analyzing the data, and making the data available to third parties who would pay for this data.

In this scenario, the third parties are business users from various lines of business and advertisers who are interested in getting subscriber analytics.

The figure below shows the innovative business model the telecommunication used. An existing analytics solution was augmented with APIs, which are provided to partners within the ecosystem, thereby monetizing resident data within the organization.

**Figure 11. Innovative business models for telecommunication companies**
The steps in this innovative business model are as follows:

1. The telecommunications company collects data from existing sources, such as call detail records that collect information for every call or text made and received by a subscriber. To increase the data’s value, it is augmented with new geolocation data sources which enable the client to associate location (latitude/longitude) for each call.
2. Data integration tools such as IBM Information Server are used to incorporate the new geolocation data with existing data in the repositories for augmented analytics.
3. All existing and enriched data is then stored in a data repository like IBM BigInsights for Apache Hadoop.
4. New analytic models are deployed in tools such as IBM SPSS to gain insight about the value of the data collected from the call detail records. Examples of data insights include identifying a caller’s daily activity (whether they go from location A to location B), establishing a network of people that a subscriber calls, and predictive modeling of when someone within a subscriber’s network would be at a location where the subscriber is.
5. This augmented data is made available to external partners through APIs with products like IBM API Management.
6. An existing data repository within the client’s organization (such as IBM DB2®) is augmented with partner authentication and usage information for billing purposes.

**Hybrid cloud deployment example**

The following table shows an example deployment of the above architecture in a hybrid cloud deployment model.

**Table 6. New business model in a hybrid cloud deployment model**

<table>
<thead>
<tr>
<th>Architectural component</th>
<th>Capability used</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services component</td>
<td>Enterprise</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Data integration</td>
<td>Client's existing data integration capabilities</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Data repositories</td>
<td>IBM BigInsights for Hadoop</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Actionable insight</td>
<td>IBM SPSS</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Transformation and connectivity</td>
<td>IBM API Management</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Enterprise data</td>
<td>IBM DB2</td>
<td>Traditional data center</td>
</tr>
</tbody>
</table>

**Scenario 4. Predictive customer intelligence and behavior-based customer insights**

In this scenario, a company uses predictive customer intelligence and behavior-based customer insight to optimize the best offers for a particular customer, to cross-sell products, and to improve customer experience.
The steps in this predictive customer intelligence scenario are as follows:

1. The data scientist develops a new process to bring social media data on her own sandbox to develop a Next Best Offer model inside of the data lake.
2. The data scientist searches the information catalog for the click-stream data, master customer data, transactions data, and interaction data inside of the data lake to find data that is necessary to develop the Next Best Offer model.
3. On her own sandbox inside the data lake, the data scientist provisions the enterprise data together with the social media data.
4. The data scientist develops new statistical functions based on the Predictive Customer Intelligence and Behavior-based Customer Insight models to create the new Next Best Offer model. She tests the results on her own sandbox.
5. The data scientist contacts the IT developer requesting to move to production the new process associated with the new Next Best Offer model. This new process consists of new data integration processes, a new NoSQL repository, a new statistical model, and a new API so all channels can access the Next Best Offer results.
6. The IT developer deploys all the components associated with the new Next Best Offer model.
7. The business analyst checks the results of the Next Best Offer model before going live for all channels.

Hybrid cloud deployment example
The following table provides an example deployment of the above architecture in a hybrid cloud deployment model.
Table 7. Predictive customer intelligence model in a hybrid cloud deployment

<table>
<thead>
<tr>
<th>Architectural component</th>
<th>Capability used</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise data</td>
<td>IBM DB2</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Data repositories</td>
<td>PureData for Operational Analytics</td>
<td>Traditional data center</td>
</tr>
<tr>
<td>Data provisioning</td>
<td>DataWorks</td>
<td>Bluemix</td>
</tr>
<tr>
<td>Actionable insight</td>
<td>IBM SPSS</td>
<td>SoftLayer</td>
</tr>
<tr>
<td>Transformation and connectivity</td>
<td>IBM API management</td>
<td>Bluemix</td>
</tr>
</tbody>
</table>

**Deployment considerations**

Cloud environments offer tremendous flexibility in the way that IT processing power is financed and delivered, often removing the need to manage and integrate many of the technologies that you need in your solutions. The need for advanced planning is reduced but still important. This section offers suggestions for better provisioning of data and computing resources.

**Primary criteria**

- Elasticity
- CPU and computation
- Data volume
- Data bandwidth
- Information governance and security

No single cloud environment optimizes all of these criteria. A little advanced planning goes a long way in deciding what type of cloud environment would be best for your company.

Figure 13 shows an optimized provisioning worksheet that balances the trade-offs between public, private, and hybrid cloud architectures. The primary criteria drive the initial architectural choice. One or more secondary criteria will tend to move the selection needle between public and private topologies.

**Figure 13. Optimized provisioning worksheet**

**Elasticity:** Elasticity is the ability for a cloud solution to provision and de-provision computing resources on demand as workloads change. Public clouds have a distinct advantage since they
generally have larger pools of resources available. You also benefit by only paying for what you use. Private clouds and dedicated hardware can make up some of the difference with higher bandwidth data paths.

**CPU and computation**: The availability of inexpensive commodity processors means the private and hybrid cloud server farms are more viable than in the past. Modern development environments using Hadoop, Spark, and Jupyter (iPython) take advantage of these massively parallel systems. Streams and high-speed analytics are an emerging area where cloud applications use more powerful processor pools to enable real-time, in-motion data solutions. Dedicated hardware allows for faster development and testing prior to migration towards hybrid and public environments.

**Data volume**: All data loses relevance over time. Data retention requires a little experimentation unless specifically governed by regulatory or other policies. Public clouds offer the flexibility to store varying amounts of data with no advance provisioning. In-house cloud storage solutions offer long-term storage cost advantages when the volume is predicted in advance.

**Data bandwidth**: Public and private clouds need to be optimized for big data. Large cloud data sets requiring fast access benefit from processing components with fast and efficient data access. In many cases, this means moving the processor to data or vice versa. Cloud systems can effectively hide the physical location of data and analytics. Tuning activities can be carried out continuously with minimal impact on deployed applications.

**Information governance and security**: As more data about people, financial transactions, and operational decisions is collected, refined, and stored, the challenges related to information governance and security increase.

Information governance policies must encompass a wider domain of data and ultimately deal with the results of related analytics that create sensitive data from inputs that are not themselves subject to safeguards. The simple fact that more people have access to data calls for better monitoring and compliance strategies. The cloud generally allows for faster deployment of new compliance and monitoring tools that encourage agile policy and compliance frameworks.

Cloud data hubs can be a good option by acting as focal points for data assembly and distribution. Tools that monitor activity and data access can actually make cloud systems more secure than standalone systems. Hybrid systems offer unique application governance features. Software can be centrally maintained in a distributed environment with data stored in-house to meet jurisdictional policies.

Public clouds are a popular choice for initial efforts. They are not the most common choice for enterprise customers. Lower bandwidth, less powerful compute environments, and governance and compliance concerns can limit the appeal of the traditional public cloud.

**The hybrid cloud**

An enterprise routinely needs a combination of public and on-premises components that, when linked, create a hybrid cloud. Generally speaking, a hybrid cloud has two or more cloud implementations with different capabilities, user interfaces, and control mechanisms.
Business that implement hybrid clouds often need flexibility and agility to deliver new capabilities. A couple of examples are:

- **Integrating social and mobile data with core business systems**: Many organizations use public cloud services to build social and mobile applications and improve the user experience. The data sources for these applications range from large social media data sets to low-latency updates based on social messaging. Linking these mobile and social systems (systems of engagement) to core business systems (systems of record) can provide greater customer insight and value. Organizations use APIs to provide access to traditional systems and data in a form that is easier to use with social and mobile applications.

- **Back-up location for disaster recovery**: Customers typically use a private cloud and switch to a public cloud in the event of a disaster to recover files. Applications and data are duplicated and synced in the public cloud. Large data sets are kept up-to-date with a mixture of continuous data transfer and smart analysis of content that minimizes bandwidth usage.

**Hybrid cloud management**

Although there are many features that make hybrid clouds appealing, there are implementation challenges. One challenge is that, by their very nature, hybrid cloud implementations involve different products and platforms.

Each platform has its own way of doing things, including tasks like:

- Configuring sets of resources, such as setting up networks or IP address pools
- Deploying new resources, such as creating a new virtual machine
- Monitoring the status of resources
- Starting and stopping virtual machines

It is difficult, even for trained administrators who work with the platforms on a daily basis, to handle the different interfaces and different capabilities. Productivity and quality can suffer as the administrators shift from product to product and are forced to change their perspectives. The challenge is even greater for casual users, ones who only occasionally need to perform routine tasks. Expecting them to master a variety of tools for different platforms is unreasonable.

The solution is to provide a "unified, single-pane-of-glass management" across the various clouds that are linked in a hybrid manner. A common, integrated administration and systems management tool that works across platforms is needed, as well as easily deployed patterns of expertise that can be used on the various cloud sites.

IBM provides a number of management capabilities to manage hybrid cloud deployments. For example, IBM Cloud Orchestrator gives access to ready-to-use patterns and content packs, helping speed configuration, provisioning, and deployment.

Bluemix continues evolving its comprehensive product catalog for deploying a plethora of services, including data services, and provides a dashboard that displays the status of currently deployed services and allows new services to be provisioned or modified.
IBM Gravitant is a consumption portal that makes hybrid and multi-cloud services easy to procure, consume, and manage. It aligns with the IT processes needed to truly make cloud work in the enterprise including planning, procurement, deployment, operations, and governance.

**Conclusion**

This article showed how IBM supports the *CSCC Cloud Customer Architecture for Big Data and Analytics* paper available on the Cloud Standards Customer Council site. This article described how to extend the architecture to a data lake and the best practices for hosting the services and components required to support hybrid, enterprise-scale analytics with cloud computing using IBM products. IBM provides first class product support for big data and analytics and the cloud architecture for customers.
Related topics

- Explore IBM Bluemix
- Designing and Operating a Data Reservoir SG24-8274-00
- *Customer Cloud Architecture for Web Application Hosting, Version 2.0* describes how IBM supports the architectural components needed to instantiate a web application hosting environment using private, public, or hybrid cloud deployment models.
- Learn more about IBM SoftLayer.
- Learn how to develop and deploy your first app on Bluemix.
- Get more information about IBM's cloud based API Management solution.

© Copyright IBM Corporation 2016
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)