Cloud
Standards
Customer
Council

Hybrid Cloud Considerations for Big Data and Analytics

July, 2017

Contents

Acknowledgements

Hybrid Cloud Considerations for Big Data and Analytics is a collaborative effort that brings together diverse customer-focused experiences and perspectives into a companion resource for IT and business leaders who are considering big data & analytics and cloud computing capabilities. The major contributors to this whitepaper are: Marcio T. Moura (IBM) and Christine Ouyang (IBM).

Executive Overview

For strategic reasons, IT leaders and their C-level business counterparts are focused on moving existing workloads to the cloud, extending existing workloads to the cloud, or building new workloads on the cloud and integrating those with existing workloads.

Quite often, the need for data security and privacy makes some organizations hesitant about migrating to the public cloud. This is perfectly understandable given the types of data gathered and used by businesses today, the regulations they must adhere to on both a local and global level, and the cost to maintain data and operational infrastructure. Fortunately, the business model for cloud services is evolving to enable more businesses to deploy a hybrid cloud, particularly in the areas of big data and analytics solutions.

A hybrid cloud is a combination of on-premises and local cloud resources integrated with one or more dedicated cloud(s) and one or more public cloud(s). The combination of on-premises and local cloud with dedicated cloud(s) is referred to as the ***"private environment."*** The public cloud and private environment are structured so that they operate independently, but communicate with each other via a secure connection on a private and/or public network, using technologies that facilitate the portability of applications and data movement.

A hybrid cloud allows organizations to integrate data from the enterprise systems on the private environment with applications running on the public cloud, while leveraging the public cloud's computational resources and storage. For example, organizations can generate actionable insights by integrating the data from Systems of Record (private environment) with Systems of Engagement in a public cloud or by applying edge-analytics on the devices in the public cloud.

In addition, hybrid cloud increases scalability by allowing organizations to use public cloud resources for situations where the private environment doesn't provide adequate computational power. Furthermore, containers can be used to increase the portability of workloads between private environments and public cloud. Hybrid cloud is the best fit for global distribution of applications and data, allowing better management of data sovereignty and compliance.

In this document, we will summarize what hybrid cloud is, explain why it is important in the context of big data and analytics, and discuss implementation considerations. This document can be used as a companion paper to the Cloud Standards Customer Council, *Cloud Customer Architecture for Big Data & Analytics* [1] to provide guidance on the deployment of big data and analytics solutions in hybrid cloud.

What is Hybrid Cloud

The ISO 17788 Cloud Computing Overview and Vocabulary standard [2] defines hybrid cloud as "a cloud deployment model using at least two different cloud deployment models" where the deployment models can be IaaS, PaaS, SaaS, and/or XaaS and the deployments involved remain unique entities but are bound together by appropriate technology that enables interoperability, data portability and application portability.

A hybrid cloud may be owned, managed, and operated by the organization itself or a third party and may exist on-premises or off-premises. Hybrid clouds represent situations where interactions between two different deployments may be needed but remained linked via appropriate technologies. As such, the boundaries set by a hybrid cloud reflect its two base deployments.

Specifically, a hybrid cloud is the connection of the private environment with one or more public cloud(s) as shown in Figure 1. It leverages the best of what each environment has to offer, providing the flexibility to locate data and services based on business need. Data can be located and accessed based on consumption patterns and analytical workload requirements within hybrid cloud environments, providing data and analytics for different personas where it is needed. Access to all areas of the hybrid cloud environments are managed and controlled to uphold privacy, security and other data governance requirements.

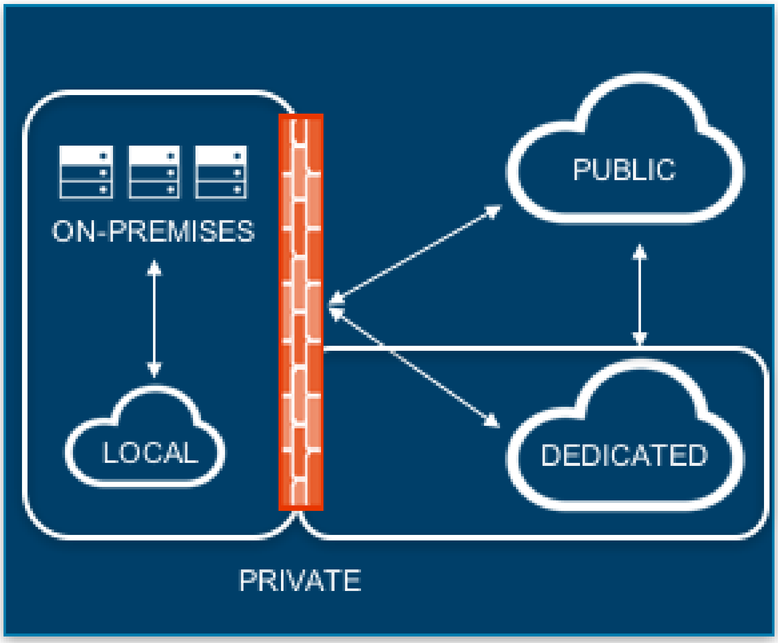


Figure 1: Hybrid Cloud Components

For many enterprises, the private environment is essential to the deployment model because most businesses will always require portions of their data and infrastructure to remain behind the corporate firewall due to industry standards, local regulations, data privacy, or their own attitudes toward controls. This creates an even more flexible architecture by giving businesses more freedom to choose and change their environments and deploy services and applications more quickly.

A hybrid cloud strategy is an overall architecture solution, and not just a migration path. The goal is to allow organizations to extend workloads from a pure in-house environment model to an extended hybrid model that couples in-house environments with public/dedicated external clouds.

Why Hybrid Cloud for Big Data and Analytics

A hybrid cloud allows different personas to work with data and analytics capabilities where it makes the most sense and helps to define the requirements where the data and analytics capabilities should be placed in the hybrid cloud environment. As a result, analytics workloads can run more efficiently wherever the data is stored.

Organizations need to consider where the data should be stored, and where the analytical processing should be located relative to the data. Therefore, consideration of external capabilities available in public/external dedicated clouds and their location, as an option, should be one of the first architectural decisions for any analytics project. Meanwhile, legal and regulatory requirements also impact where data can be located, as many countries have data sovereignty laws that prevent data about individuals, finances and intellectual property from moving across country borders. More detailed information on data residency can be found in the CSCC's *Data Residency Challenges* whitepaper. [3]

Systems are going to have multiple centers of gravity, which will dictate where processing will occur. For example, if building a data lake as part of a Systems of Insight solution and the data that feeds the lake is in the private environment, the center of gravity will within the private environment and processing on that data should stay within the private environment. But if the Systems of Insight solution includes data born on the cloud, there could be a second center of gravity.

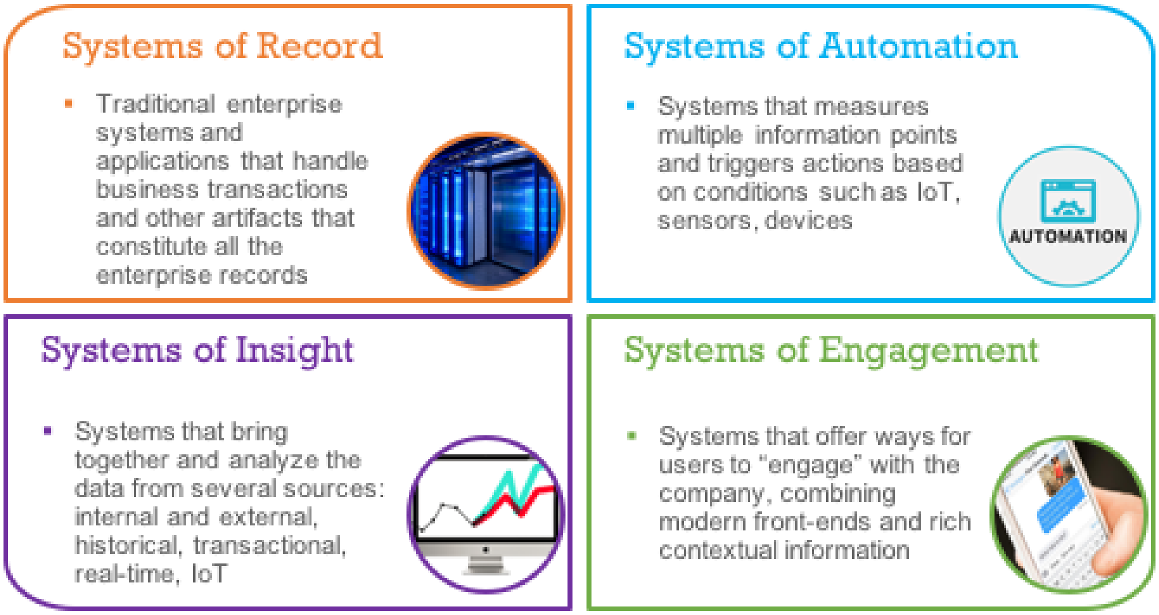
Primary Drivers for Hybrid Cloud

- **Integration:** Organizations need to integrate data that is stored and managed in a hybrid environment across traditional IT (on premises)/private cloud environments with public cloud services. Typically, these organizations need to integrate Systems of Engagement and/or Systems of Automation applications, such as social media, customer management systems, and devices with Systems of Insight, such as predictive and real-time analytics hosted on public or private clouds, and mission-critical applications and data stored on servers in on-premises data centers (Systems of Record).
- **Brokerage/management for workload and resource optimization:** Different workloads have different requirements for security, speed, resources and storage. Many organizations are driven to hybrid cloud because they want the option to place data and analytical workloads where it makes the most sense based on business requirements. These organizations want the ability to optimize cost, performance and agility, while also enjoying the flexibility to move data and analytical workloads between on-premises environments, private cloud environments and public cloud.

- **Portability:** Another major case for hybrid cloud is the need to ensure portability of analytical workloads and data. In order to manage costs and effectiveness, IT management needs to be able to move workloads and data to whatever platform best meets changing customer demands. This capability requires IT to consider the feasibility of new analytical workloads and data on a specific hybrid cloud environment based on the overall hybrid cloud architecture.
- **Compliance:** A hybrid cloud allows for distributing global applications, data and workloads across geographically dispersed environments: on-premises, private cloud and public cloud; where requirements for data sovereignty, compliance, privacy, identity management, and data protection could dictate that data and consequently workloads be placed on a specific environment in a specific country.

Digital transformation requires a hybrid cloud that is open and flexible by design, and gives clients the freedom to choose and change environments, data, and services as needed. This approach allows cloud applications and services to be rapidly composed using the best relevant data and insights available, while maintaining clear visibility, integrated control, governance, and security everywhere.

Figure 2 highlights the enterprise IT domain systems that play an important role in the hybrid cloud strategy where each system can be part of one or more environment(s) of the hybrid cloud.



Systems of Record

- Traditional enterprise systems and applications that handle business transactions and other artifacts that constitute all the enterprise records

Systems of Automation

- Systems that measures multiple information points and triggers actions based on conditions such as IoT, sensors, devices

Systems of Insight

- Systems that bring together and analyze the data from several sources: internal and external, historical, transactional, real-time, IoT

Systems of Engagement

- Systems that offer ways for users to "engage" with the company, combining modern front-ends and rich contextual information

Figure 2: Enterprise IT Domain Systems

As shown in Figure 3, the majority of the Systems of Records usually resides in the on-premises environment, while the Systems of Engagement and Systems of Automation are mostly deployed in the public cloud, and the Systems of Insight span across all hybrid cloud environments. The flexibility and openness of the hybrid cloud allows the data and the associated analytical workload to be placed where it makes the most sense in terms of business needs. The information privacy and security is managed and controlled consistently across all the systems of the hybrid cloud environment.

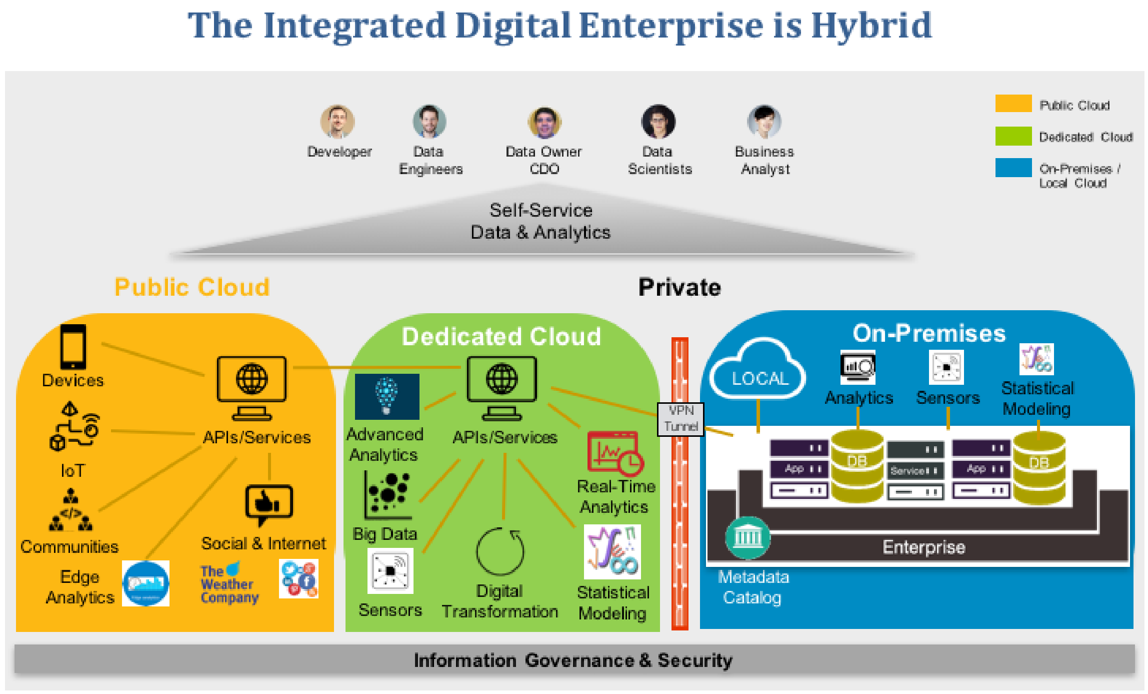


Figure 3: Hybrid Cloud for the Digital Enterprise

The use case that is one of the primary drivers of hybrid cloud is the integration of Systems of Engagement (SOE) and/or Systems of Automation (SOA) with Systems of Insights (SOI) and Systems of Records (SOR). This is the primary case for extending private data to the public cloud. Organizations want the full agility and flexibility to generate personalized customer offers and respond to market dynamics. This use case requires a new type of architecture – a secure enterprise *data lake* – a group of fit-for-purpose repositories that are well managed, governed, protected, connected by metadata and provide self-service access. How does a typical enterprise data lake fit in the hybrid cloud strategy landscape, including different IT domains such as Systems of Record, Systems of Engagement, Systems of Automation, and Systems of Insight? It depends on the data gravity of the organization. If the data gravity stays inside of the private environment, then the data lake will be defined inside of the private environment. If the data gravity is shared between the private environment and public cloud, then the data lake will be defined across both the private environment and public cloud. If the data gravity is within the public cloud, then the data lake will be defined on the public cloud.

Deployment Considerations

A general guide of hybrid cloud deployment considerations can be found in the CSCC's *Practical Guide to Hybrid Cloud Computing* [4]. Some of the key topics that need to be highlighted when planning a hybrid cloud strategy for big data and analytics are:

- **Data and analytics** – where to store the data based on the analytical workload and the type of access? What type of data is stored in each environment?
- **Data movement and replication** – where to store the data to avoid data movement and replication? How is data synchronized between the systems?
- **Data preparation and integration** – what is the best environment for data preparation/integration? Where is the data processed?
- **Data sovereignty and compliance** – where should data be stored based on data sovereignty requirements and which compliance regulations are required?
- **Data governance and security** – how to secure/protect the data across all environments and ensure appropriate access control is in place? How is metadata handled?
- **High availability and disaster recovery** – how is high availability and disaster recovery provided for the application/data in the cloud?
- **Network configuration and latency** – what network configuration is appropriate to satisfy the latency requirements of an application in the cloud?
- **Workload and portability** - does the application and associated data need to be portable to different environments?
- **Scalability** - does the application require the capability to scale up and/or down?
- **Resource orchestration** - how to define the resource orchestration required for all workloads?

These aspects are discussed in more detail below.

Data and Analytics

How does an organization decide where to put data on a hybrid cloud and how to use it? What's the best strategy to balance sharing and mobility with the need for privacy and security in a hybrid cloud? Data can be a valuable asset and/or form of intellectual property, so organizations are naturally concerned about moving it. Across industries, organizations are still trying to figure out the best use cases to move to the cloud, and most have not started a hybrid cloud strategy yet. In terms of big data, often the lack of clarity on data classification and privacy is the first challenge to be addressed for the organization to define where to put the data in the hybrid cloud.

One of the most important hybrid cloud considerations for all organizations is data gravity. To limit data movement, production workloads should be processed where the data is stored. This means that an organization should consider the analytical workloads that they will be processing when deciding where to locate their data. Another important consideration is where data discovery and exploration, and advanced analytics will happen when deciding where to locate the data.

A well-managed hybrid cloud data lake helps organizations overcome data silos by providing the data and metadata with information governance based on business needs allowing users to access the data they need from where it makes sense. This type of on-demand data engagement produces insights when they are needed most. Organizations should strive to get as much data into the data lake as possible, while implementing information governance policies, standards and tools to control user access and data provenance. A well-managed data lake also allows self-service capabilities and multispeed IT, enabling users such as data scientists and business analysts to locate data by using the information catalog to easily conduct analytics without IT assistance to generate their own insights. The data lake can be located in the private environment or can also be located in the private environment and public cloud together. The best way to define where the data lake should be located is to consider the key topics when doing a data topology exercise as described above.

Organizations are using the cloud for a range of business use cases, including reporting, sandboxes, production, IoT, analytics, and much more. Cloud analytics services offerings are evolving and becoming more popular, especially with business customers. Having a hybrid cloud architecture that can provide access to new technologies as they emerge without requiring IT departments to learn, install or support them can be a significant accelerator for business analytics solutions. Today's new SaaS offerings are increasingly targeting specific business areas such as churn detection as a service, fraud detection as a service, and marketing campaign as a service. These offerings can improve business outcomes much faster than in-house efforts. They also require the integration of private environments with public cloud - another driver for a hybrid cloud strategy.

The information flow of the data analytics lifecycle on hybrid cloud is composed of three distinct steps where any step can be done on the private environment and/or public cloud:

1. Search and gather information using the information catalog (e.g., inventory of data assets) and the data from all hybrid cloud environments.
2. Generate business insights (analytics) using data from all environments of the hybrid cloud.
3. Deploy those insights into the business process on any hybrid cloud environment where it makes the most sense before repeating the cycle.

The hybrid cloud allows different personas to work on different environments and data topology zones, collaborating with each other to generate new insights. A reference architecture for building big data and analytics in the cloud that supports the analytics lifecycle is described in the CSCC's whitepaper *Cloud Customer Architecture for Big Data and Analytics* [1].

Data Movement and Replication

One of the biggest challenges of hybrid cloud is the movement of data within the hybrid environments. By its very nature, data can be cumbersome and difficult to move. In addition, data needs to be carefully protected and controlled as it is being moved. Copying large volumes of data to the public cloud currently takes too much time and bandwidth - doing it incrementally with micro-batches, streaming, or asynchronous replication can be an option to reduce time and latency. Today, all cloud providers require copying data to a disk appliance to move the data to the cloud if the volume is greater than 5 TB.

Data synchronization is another challenge. When changes are made on the prime copy (private environment), what is the best approach to replicate the changes to the public cloud instances to prevent out-of-sync information?

The current options for data movement among the hybrid environments includes the use of disk appliances for large data volumes, the use of traditional file transfer protocols for small data volumes, and high-speed file transfer tools for medium/large data volumes.

Examples are:

- Import/Export using a disk appliance/USB drive for large data volumes; data can be compressed and encrypted.
- FTP/SFTP using TCP protocol for small data volumes; can be susceptible to latency.
- REST API calls or messaging API via Kafka for small data volumes; can be secured with HTTPS, and can be susceptible to latency.
- Tsunami UDP using a file transfer protocol including TCP for control and UDP for data to transfer small/medium data volumes over very high speed long distance networks.
- Parallel Remote File Transfers using bbSCP, bbFTP, or gridFTP for small/medium data; data can be compressed but it cannot be encrypted.
- Data replication using WANDisco for medium/large data volumes; can be scaled out to maximize bandwidth, and can be used for DR replication.
- High-speed file transfer using Aspera for medium/large data volumes; supports compression, encryption, and auto scaling.

The current options for moving data within hybrid cloud suffer from the low speeds and high latency provided by typical wide area networks (WAN). Latency in these environments can reach into the milliseconds, compared with microseconds expected in the private environment. This means public cloud instances may be inefficient and slow. Therefore, latency is a very important factor when considering different scenarios on a hybrid cloud such as:

- **Operational applications on the cloud** – the data should be collocated with the application to avoid latency.
- **Large volume data movement** – should avoid large scale data movement even with VPN and compression.
- **Data movement** – the best option is continuous data transfer to move data to the cloud.

There are a few possible ways to reduce/improve latency in hybrid cloud environments. The most useful option to reduce WAN latency is to have a dedicated/private network connecting the public cloud to the private cloud and/or on-premises environment.

The top four ways to reduce latency and maintain performance:

1. **Location:** It takes more time to cover longer distances, even when moving at the speed of light. To reduce hybrid cloud latency, organizations should connect to a public cloud facility that is geographically close to its private environment.

2. **Dedicated connections:** The Internet is a single network that is shared amongst billions of users. This means public Internet connections may experience traffic congestion, causing bottlenecks and increased latency between private environments and public cloud(s). To reduce hybrid cloud latency, organizations can establish dedicated private connections between their private environment and public cloud(s).

3. **Optimize traffic:** Even with caching, there are instances when data has to be moved within a hybrid cloud, and this can cause an increase in network traffic. Latency will occur each time a packet moves across the Internet, but applications can reduce the cumulative latency of those packets by moving fewer of them. Data compression allows an organization to pack more data into each packet, which lowers the number of packets needed to move a data set.

4. **Optimize workloads:** Developers should not overlook the importance of workload design and its influence on network latency. Workloads designed to use massive data sets, or deal with critical, time-sensitive data, can be extremely sensitive to network latency. Architecting these workloads to better accommodate hybrid cloud environments can alleviate some latency issues.

Data Preparation and Integration

Data preparation is not just for data professionals. What separates data preparation from more traditional data management activities is really nothing more than the level of technical knowledge that is required. While data management is appropriate for database administrators, IT managers, and others who are close to the data in their everyday jobs, data preparation is more approachable for a less technical audience. Marketing managers, sales managers, and financial decision makers are taking an active role in improving the quality of their data, and thus taking more control over the ultimate quality of their decisions using a hybrid cloud environment.

Effective analytics relies heavily on data preparation. The old expression “garbage in, garbage out” still applies, but today’s organizations are going beyond the basics to improve data quality by empowering business users to perform self-service data prep and by providing new tools on the cloud to perform those tasks. Even the highest quality data, in massive volumes, can lead to tremendous amounts of time wasted in searching, gathering, integrating and preparing data for analysis. Best-in-class organizations are intensely focused on using technologies and activities aimed at enhancing both the quality and the timeliness of their data in order to improve their analytics. To accomplish that requires an environment that allows self-service and collaboration with access to all data from the organization. A hybrid cloud strategy helps organizations to define the environment for their business to improve the quality and the timeliness of the data, and the use of off-the-shelf tooling.

The three phases of achieving hybrid cloud data integration are:

1. **Exposing private data to SaaS applications:** The first stage in developing a hybrid integration platform is to expose Systems of Record and Systems of Insight data to SaaS applications on the cloud via API management and a secure gateway. For example: a financial institution wants to monetize data by exposing data and insights to customers and third party vendors from their systems of records and their systems of insights (on their private environment) to the systems of engagement (in the public cloud).

2. **Hybrid cloud data lake:** As the volume and variety of data increases, enterprises need to have a data topology strategy based on data tiers, consumption patterns, analytical workloads, and data gravity to define the different information zones of the hybrid cloud data lake including landing zone, provisioning zone, shared operational zone, data warehouse and data marts zone, self-service and collaboration zone, and analytics zone. The data topology strategy will define where the data should be stored and processed on the hybrid cloud data lake taking into consideration the type of consumptions, the type of analytical workloads, the type of data integration among data repositories, the type of data movement required, the type of data sovereignty and compliance required, and the type of data governance and data security required.

3. **Real-time analytics with streaming data:** Businesses today need insight at their fingertips in real time. In order to prosper from the benefits of real-time analytics, they need an infrastructure that can support it. As big data analytics and IoT data processing moves to the cloud, companies require fast, scalable, elastic, and secure platforms to transform that data into real-time insights. This is being accomplished with edge-analytics running in the cloud consuming the information from IoT devices. The information is also sent to a data repository in the hybrid data lake for further discovery and exploration.

Data Sovereignty and Compliance

One of the basic tenets of cloud computing is the ability to provide access to resources across a geographically dispersed cloud environment. This makes the cloud ideal for global distribution of applications and data. But, how should geographies that have highly restrictive data sovereignty laws or practices, such as Europe and Asia, be handled?

The data sovereignty solution for hybrid cloud should include the following components:

1. **Security standards –** The solution requires a strong set of security standards applicable to private environments and public cloud. It is important when developing a hybrid cloud solution to extend security standards from the private environment, as much as possible, to the public cloud.

2. **Data loss prevention (DLP) monitoring and controls –** DLP solutions define controls over the flow of data and monitor that data flow to detect data breaches.

3. **Data aware services –** As services are developed, the integrated software components need to have proper authorization and filtering capabilities. For example, an identity management system where the directory services replicate data between geographically dispersed instances, based on filtering rules defined by data governance.

4. **Data segmentation (DS) across a hybrid cloud infrastructure** – Countries or organizations may require different levels of DS control requiring that data must be maintained in a specific location. In this case, a straightforward solution comes in the form of hybrid cloud with regional instances located at or in proximity to the point of high DS requirement.

5. **Consistent data management tools** – Common and consistent data management tools and practices across all cloud regions with privacy and security controlling who is authorized to administer a given instance, or data set.

Maintaining and demonstrating compliance can be more difficult with a hybrid cloud. Not only do customers have to ensure that their public cloud provider and private environment are in compliance, but they must also demonstrate that the means of coordination between the two environments are also compliant. Also, the cloud providers have the responsibility to provide environments and services that are compliant with the industry and government regulations.

The data compliance solution for hybrid should have the following components:

- **Compliance –** Comply with industry and government regulations including SOX, HIPAA, ITAR, SOC2, and PCI.
- **Encryption –** All data is encrypted at-rest and in-motion with the highest levels of encryption.
- **Full Auditing –** Every user action in the hybrid cloud environment is logged, and audit reports can be easily created to provide full visibility and demonstration of compliance.
- **Data Sovereignty –** Geographic policies can be enforced to restrict user content to storage physically located in specific countries to meet data residency requirements.
- **Secure Containers –** All content is protected with encryption and access controls in all data repositories on the hybrid cloud.

Data Governance and Security

Good hybrid cloud data governance implies several priorities for IT and the business:

1. **Broad agreement on what information means**, including metadata on common policies and plain-language rules for the information the business needs and how it will be handled.

2. **Clear agreement on who owns the information assets, and how owned information assets will be maintained and monitored**, for example, operational data quality rules for master data management in on-premises systems.
3. **Enterprise and departmental standard practices for securing and protecting strategic information assets**, such as defining who owns the information, defining role-based access to information, creating rules governing how information is shared, and protecting sensitive information from third parties.
4. **Enterprise data integration strategy that includes lifecycle management**, architecting how data will flow and be assembled into strategic information, and also understanding how that data/information will be maintained over time including archival and defensible disposal.

Security continues to be the primary concern surrounding public cloud adoption, and these challenges also apply to hybrid cloud. As organizations consider which application can run where, they need to consider the compliance, identity management, and data protection needs of those application workloads. Also, organizations need to consider the following security aspects:

- Is a virtual private network required?
- Does data in motion and at rest need to be encrypted?
- Are secure gateways required?
- Is a secure communication protocol (HTTPS) required?
- Is a secure encrypted tunnel (SSH Tunnel) required?
- Is LDAP authentication required?
- Where do applications and data need to be located?
- How is security monitored across hybrid cloud services?
- How are on-premises security processes and solutions adapted in hybrid cloud?
- Who manages the security (e.g., cloud firewall)?

Privacy regulations may limit certain workloads from crossing geographical boundaries. In addition, regulatory requirements such as HIPAA, PCI and SOX require the infrastructure where data resides for a specific application to be compliant.

For identity management and credentials, operations teams need to ensure user permissions and unique credentials propagate from private environments to public cloud environments. Lastly, customers must ensure that public cloud providers have the basic data protection and cryptographic mechanisms in place and are diligent about updates and patches with minimal disruptions.

Mind the gap: The biggest potential point of failure for hybrid cloud deployment is where the public cloud and private environment offerings meet. At this point, a gap often exists between in-house security protocols and third-party security standards. If this gap is not closed, malicious actors or malware could slip through it. Meeting this challenge requires a new breed of IT professional, one who is familiar with both the rigors of in-house penetration testing and the more flexible nature of public cloud environments. With the right amount of oversight, it is possible to close this gap and improve hybrid cloud security at its weakest point.

Deal with data: Data is the most valuable resource a company owns. Yet in public and hybrid cloud deployments, the security of data is often overlooked in favor of ease of access and usability considerations. This, of course, leads to an increased security risk. To meet this challenge, companies must define a specific data handling and encryption strategy including encryption key management, encryption for data in motion and at rest, secure gateways, virtual private networks, and data location before deploying cloud services. This eliminates the problem of ad hoc data security — which, by nature, is reactive rather than proactive — and replaces it with reliable, repeatable security protocols that can be applied both cloud-wide and company-wide. Furthermore, organizations need efficient data governance, metadata, and classification strategies to ensure that data has the correct classification anywhere on the hybrid cloud, and the data access granted to different users is based on the classification of the data.

Get compliant: Compliance is a high-profile buzzword across the tech industry. That's because if cloud deployments are not compliant with industry or government regulations, companies could face monetary fines and sanctions. Meeting this challenge requires a shift in local IT focus away from pure technology management toward a larger-scale view that's focused on ensuring compliance across the hybrid cloud environments. Organizations need to understand the compliance level that an application requires in a specific environment of the hybrid cloud. The compliance level is dependent on privacy regulations associated with data sovereignty rules not allowing data to cross geographical boundaries, and regulatory requirements such as HIPAA, PCI and SOX. The different compliance levels are easily enforced within a private environment but it is more difficult to enforce on the public cloud where only the cloud providers have control.

A hybrid cloud security framework must consider the four components described below.

- **Manage access:** Role-based, automated identity and access systems must bridge into the public cloud and back into the private environment to help ensure that the right people have access to the right information. Furthermore, different kinds of access have different risk profiles and must be authenticated and monitored accordingly.
- **Protect data:** Data volumes have exploded concurrently with cloud adoption. The long list of credit card and personal information breaches is proof of how valuable data can be to groups of criminals. It is important to extend the data security infrastructure to protect information regardless of the data source and the data format (including voice, image, video, sensor data, and other types of traditional formats) or where the data is stored (private environment, public cloud, or hybrid). Many data breaches are the result of preventable vulnerabilities in applications that access the data. Application security scanners available on the development platform itself allow developers easy access to security testing and remediation services that

they can build into their agile workflow and help reduce vulnerabilities to protect against application-based data breaches.

- **Gain visibility:** Many security solutions offer visibility into only a portion of security risks. An enterprise may have embraced a large number of these vertical solutions to help protect against myriad threats. Adoption of the cloud increases the number of security solutions required for monitoring and enforcing IT security further obscuring threats with large amounts of disparate information. Instead, enterprises should adopt security intelligence solutions to evaluate contextual clues and provide a single dashboard to view activity and threats across private and public cloud environments.
- **Optimize security operations:** Security environments grow complex over time as new technologies are adopted and the demand for protection against a broad range of threats increases. Seek out expertise required to assess security practices. Plan, design, and build out world-class security operations centers to extend security to the cloud.

Refer to the CSCC's whitepaper *Security for Cloud Computing: 10 Steps to Ensure Success* for details on cloud security requirements. [5]

High Availability and Disaster Recovery

Historically, Disaster Recovery (DR) solutions have been expensive and complex to deploy. Only the largest organizations can afford DR – and only if they had a technically sophisticated IT team. The issue of complexity takes its toll on many teams who don't anticipate the extensive configuration and setup necessary to maintain the system, the complex mechanisms for DR drills, or the challenges with scale.

With adoption of private environments, DR solutions have started to leverage the benefits of virtualization (compute and networking) and enabled DR drills against replicas running on a recovery site in an isolated network. This capability allows IT administrators responsible for DR to perform DR drills periodically without the fear of impacting production.

Even with virtualization, however, many of the other challenges around complexity persist. For example, a complete end-to-end DR solution should offer DR capabilities beyond just data replication, e.g., bringing reliability to apps on the recovery site with optimized recovery time objectives, or ensuring connectivity to clients through required network configuration and compliance needs like reporting. Simple virtualization doesn't resolve the need for high-friction maintenance of multiple components. Other concerns include the deployment and ongoing maintenance needed to ensure high availability of the DR solution.

Managing the DR strategy in a hybrid cloud environment has the following advantages:

- *Simplicity*. The simplicity of a hybrid environment ensures straight forward connectivity/switching between the DR protocols in a primary site vs. secondary backup sites.
- *Reduced cost*. The flexibility and ability to personalize a hybrid environment means that a pay-as-you-go model can be leveraged to start small and expand the DR protection needs gradually.

Skipping the need for a big DR budget commitment up front can go a long way to help the organization get started.

- *Scalability*. A hybrid cloud can provide a scalable solution which can meet any organization's need for DR solution.
- *Regulatory and compliance readiness*. Hybrid cloud enables compliance with regulatory and compliance norms as applications are deployed in on-premises datacenters and the application's data is replicated and encrypted on the private network to a recovery site.
- *High availability*. A highly available cloud platform can provide protection from both local and regional disasters. For example, if there are compute or storage failures, the service continues to operate normally because of the high availability provided by the hybrid cloud compute and storage platforms which are replicating multiple copies.
- *Access from anywhere*. It's important to have the DR service deployed in a hybrid cloud that can be accessed from anywhere.

A holistic approach to high availability is focused on the two key elements of any cloud environment: technology and processes. Defining a holistic approach strategy in a hybrid cloud environment has the following advantages:

- *Improved operational availability*. Near-continuous application, data, and system availability across geographically dispersed cloud environments.
- *Reduced risk*. Understand, manage, and address potential threats to business continuity, ranging from natural disasters to facility disruptions to technology equipment failure.
- *Better recovery*. Maintain data consistency by mirroring critical data between primary and recovery sites in a hybrid cloud environment.
- *Cost management*. Invest strategically in protecting those critical data and application assets that need it most for the organization.

Network Configuration and Latency

A hybrid cloud requires a thoughtful network design and considerations for multi-tier applications. The impact of latency between the public cloud location(s) and the private environment may not be something the end-user or the application tiers are willing to tolerate.

The network bandwidth must also be considered as a key cost driver or impediment to migrating and managing workloads. Appropriate bandwidth is required for transferring large data sets and should be taken into account when deciding what is burst worthy. In addition, most public cloud providers do not charge an organization for uploading data (unless they are leveraging a direct connection) but will charge for downloading the data back, which is required in a burst scenario.

In terms of configuration, the IP blocks that have been assigned for the network topology may need to be reconsidered in a hybrid scenario. In addition, the traffic network policies connecting the various tiers of a multi-tier application and the control mechanism for routing traffic between the tiers will also have to be evaluated. Lastly, network security policies (e.g., VPN setup), firewalls and encryption of the communication flow need to be extended from the private environment to the public cloud.

Is Software-Defined Networking (SDN) a Solution to Hybrid Cloud Networks?

The reality of networking for hybrid clouds today is that elements in many disparate domains must be stitched together, and they are only able to deliver an approximation of the responsive, elastic platform that customers require. A usable network service combines network elements in several key areas: IP routing; wireless and wireline access, metro and regional aggregation; packet and optical cores and other domains. In most cases, this involves integrating multiple vendors' platforms. But under current implementations, operators are constrained in their ability to support a variety of services with the elasticity and responsiveness required in hybrid cloud computing. To overcome these complexities, service providers' cloud networks need to evolve and become as responsive and nimble as customers need. SDN may be what operators need to meet some of the challenges inherent in hybrid cloud networking. It accomplishes this using a design similar to the one that's made cloud computing possible - abstracting software from underlying hardware and unleashing the power of both.

Architecturally, SDN allows network services to be abstractly defined once and "overlaid" broadly onto a population of heterogeneous network elements. By creating this consistently-defined middle tier of network services, operators can simplify and expedite the processes for introducing new offerings. Consequently, providers can overcome some of the key hurdles they face in creating attractive services for hybrid cloud computing customers. They can use SDN to implement a more straightforward method for creating services with latency, capacity, resiliency, security or time-to-activate guarantees tuned to the demands of hybrid cloud computing customers.

When selecting a cloud service provider, customers must consider the provider’s SDN strategy and current implementation, specifically in a big data and analytics context.

Workload and Portability

Workload complexities are more prevalent in hybrid cloud environments than in single cloud architectures. Some workloads may be permanent and need to run constantly, such as an online commerce site or a control system that manages a critical environmental process. Virtualized workloads (i.e., workloads deployed via virtual machines) add another level of complexity. Business services and various application models are also added into the mix. In a hybrid cloud environment, workloads may be running on different environments, running different kinds of infrastructures using different operating systems that often have to behave as though they are a unified system.

Operationalizing Workloads

What is the connection between workloads and workload management in the cloud? It is actually at the center of determining whether it is a well-performing hybrid cloud environment or not.

When looking at workloads from an operational perspective, it becomes clear that lots of issues need to be taken into account when determining how to create an overall hybrid cloud environment that both performs at a quality level and meets security and governance requirements. This is not a static requirement; from an operational perspective, organizations need to be able to dynamically change workload management based on changing business requirements.

APIs: Key to Managing Workloads in the Cloud

Application programming interfaces (APIs) enable a software product or service to communicate with another product or service. The API specifies how one application can work together with another. It provides the rules and the interfaces. The developer doesn't need to know the nitty-gritty of the application because the API abstracts the way these programs can work together. An API also provides an abstracted way to exchange data and services. APIs are important for managing workloads in a cloud environment. Because of this abstraction, the API can hide things from developers. For example, an outside developer doesn't need to know the details of the internal security, so those details of the system are hidden. The API allows the developer to execute only the intended task. Another important point is the API Economy where companies (providers) expose their internal business assets or services in the form of APIs to third parties (developers/consumers) with the goal of unlocking additional business value through the creation of new assets.

The Necessity of a Standard Workload Layer

No standard API exists that allows developers to work with different cloud models provided by different cloud vendors. What is actually needed is a standard layer that creates compatibility amongst cloud workloads. In service orientation, the XML model allows for interoperability amongst business services. There is no equivalent model for hybrid cloud. This standard workload layer requires the definition of standard software / tools / interfaces / communication protocols available across all cloud providers. Cloud providers need to collaborate with each other to define the standard workload layer.

Workload Portability

Discussing APIs and standards is essential because workload management is fundamental to the operation of hybrid cloud. In a hybrid cloud environment, being able to move workloads around and optimize them based on the business problem being addressed is critical. Despite the fact that workloads are abstracted, they are built with middleware and operating systems. Workloads must be tuned to perform well in a specific hardware environment. In today's hybrid computing world, a lot of manual intervention is needed to achieve workload portability. A hybrid service workload broker needs to be created to provide a layer that can examine the infrastructure of the underlying cloud-based service and provide a consistent and predictable way to handle different workloads as if they were built the same way.

Cloud Foundry, OpenStack, and Docker are gaining traction as the preferred approach to ensuring portability across clouds. Container technologies are used to create an encapsulated environment for applications making it simpler to deploy the application in either public or private clouds or even on-premises.

Scalability

The ability to scale on demand is one of the biggest advantages of cloud computing. Often, when considering the range of benefits of cloud, it is difficult to conceptualize the power of scaling on-demand, but organizations of all kinds enjoy tremendous benefits when they correctly implement auto scaling. Many of the issues and challenges experienced before the advent of cloud are no longer issues. Engineers now working on cloud implementations remember working at companies that feared the Slashdot effect – a massive influx of traffic that would cause servers to fail.

With auto scaling, the risks associated with traffic overflow causing server failure can be greatly reduced. Furthermore, and somewhat contrary to our intuition, auto scaling can reduce costs as well. Instead of running instances based on projected (assumed) usage and leaving excess resources in place as a buffer, resources matched to actual usage are run on a moment-to-moment basis.

The price and scalability advantages are not without their own complexities. While the environment can be scaled on demand, applications need to be able to scale with the environment. This might seem straightforward when running a website benefitting from an elastic load balancer distributing traffic across multiple instances that scales with increased demand. Yet, there are other considerations that need to be made when accounting for scaling session information, uploads, and data.

Compared to legacy IT management, the most important paradigm shift in cloud computing is that cloud systems are transitory, and anything on them needs to be completely and immediately replaceable. For instance, rather than storing data locally, use a cloud storage solution. If the business cannot move systems and data onto cloud storage, a distributed file system may need to be considered. Session information should no longer use local file stores; rather consider using in-memory cache services to save the sessions.

The issues around scalability are not new problems. Now, with auto scaling, the systems being scaled simply become CPU and memory, and developers write data to a long-term store. There are many ways to get a system from zero to 100% scalability. But no matter what approach is being used, the ability to scale is limited only by the ability of the application to scale with it.

Resource Orchestration

The diversity of Cloud Resource Orchestration Frameworks (CROFs) make the decision making process hard for software engineers, solution architects, or administrators who want to migrate their applications to cloud. Having concrete dimensions that give insight into comparative features of CROFs eases the task of cloud framework selection. Consider the following dimensions:

- *Application Domain.* This dimension refers to the type of applications that the frameworks have been targeted and customized for including multi-tier web and Content Delivery Network (CDN) applications, and large-scale data processing. Multi-tier web application refers to migrating in-house web, mobile, or gaming applications to public cloud or private environments in order to meet scalability and availability requirements. Development and test activities are conducted in cloud environments.
- *Resource Type.* This dimension refers to the resource type: infrastructure or hardware resources such as network, CPU, and storage; platform resources including application servers, monitoring services, database servers, etc.; software components and sub-processes such as SaaS offers.
- *Resource Access Component*. This dimension refers to the mode of interaction with the cloud resources. Interfaces are the high-level abstractions through which an administrator manipulates cloud resources. Currently, there are three types of interfaces supported by resource orchestration frameworks: low-level command line tools that wrap the cloud specific API actions as commands or scripts; web based system dashboards that represents cloud resources through user-friendly, visual artifacts, and resource catalogs; web services APIs that enable other tools (e.g. monitoring tools) and systems (e.g. provisioning systems) to integrate or use cloud resource management operations into their functionalities.
- *Interoperability*. One of the key barriers for cloud computing adoption is interoperability. Therefore, platforms that have higher levels of interoperability with other public cloud or private environments have a higher chance of customer adoption. Consequently, this dimension refers to the ability of a resource orchestration framework to port applications across multiple environments or to use resources from multiple environments of the hybrid cloud for composing and hosting applications. Interoperability is necessary to avoid cloud provider lock-in. However, designing and implementing generic resource orchestrators that can work with hybrid cloud is nontrivial as it requires APIs specific to each environment.
- *Resource Selection*. This dimension refers to the level of automation supported by an orchestration framework with regards to selection of software and hardware resources. The selection process involves identification and analysis of alternatives (cloud resources) based on the preferences of the decision maker (administrator). Making a selection implies that there are alternative choices to be considered, and in such a case, administrators not only need to identify as many of the alternatives as possible but also to choose the one that best fits their selection criteria.
- *Run-time Adaptation.* This dimension refers to the degree to which a resource orchestrator is able to adapt to dynamic exceptions. Adaptation, in general is realized either manually or automatically. In a manual manner, provided an event occurs (for example, reaching a threshold) the framework does not provide any auto-scaling facility and in the best case it will alert the administrator via an email or text message to manually configure the instances in order to accommodate to new conditions. On the contrary, in an automatic fashion the frameworks

will adapt to exceptions through reactive and predictive techniques. Reactive techniques respond to events only after reaching a predefined threshold that is determined through monitoring the state of hardware/software resources.

Deployment Scenario

The reference architecture and components of a big data analytics solution in the cloud are described in the CSCC's whitepaper *Cloud Customer Architecture for Big Data and Analytics* [1]. Now that we have described the considerations of hybrid cloud for big data and analytics, let's look at how the solution components are implemented in an example hybrid cloud scenario using the reference architecture.

Cyber Threat Intelligence

Clients are looking for new ways to combat cybercrime in real time, especially since 100,000 new malware items are introduced every week. Signature learning methods are not enough and a scalable architecture with machine learning capabilities is required to analyze data in real time and in batch to detect advanced persistent threats.

Business Challenges

Some of the significant challenges that are unique to the cyber threat architecture implementation include:

- Improve situational awareness of network security from both external and internal sources.
- Introduce a new service in the marketplace, which capitalizes on leveraging the network. backbone provided by the telecom provider and analytics to provide per-client and per-industry threat analysis.
- Monetize their investment in a cyber threat intelligence solution by delivering it as a managed service to their enterprise clients.

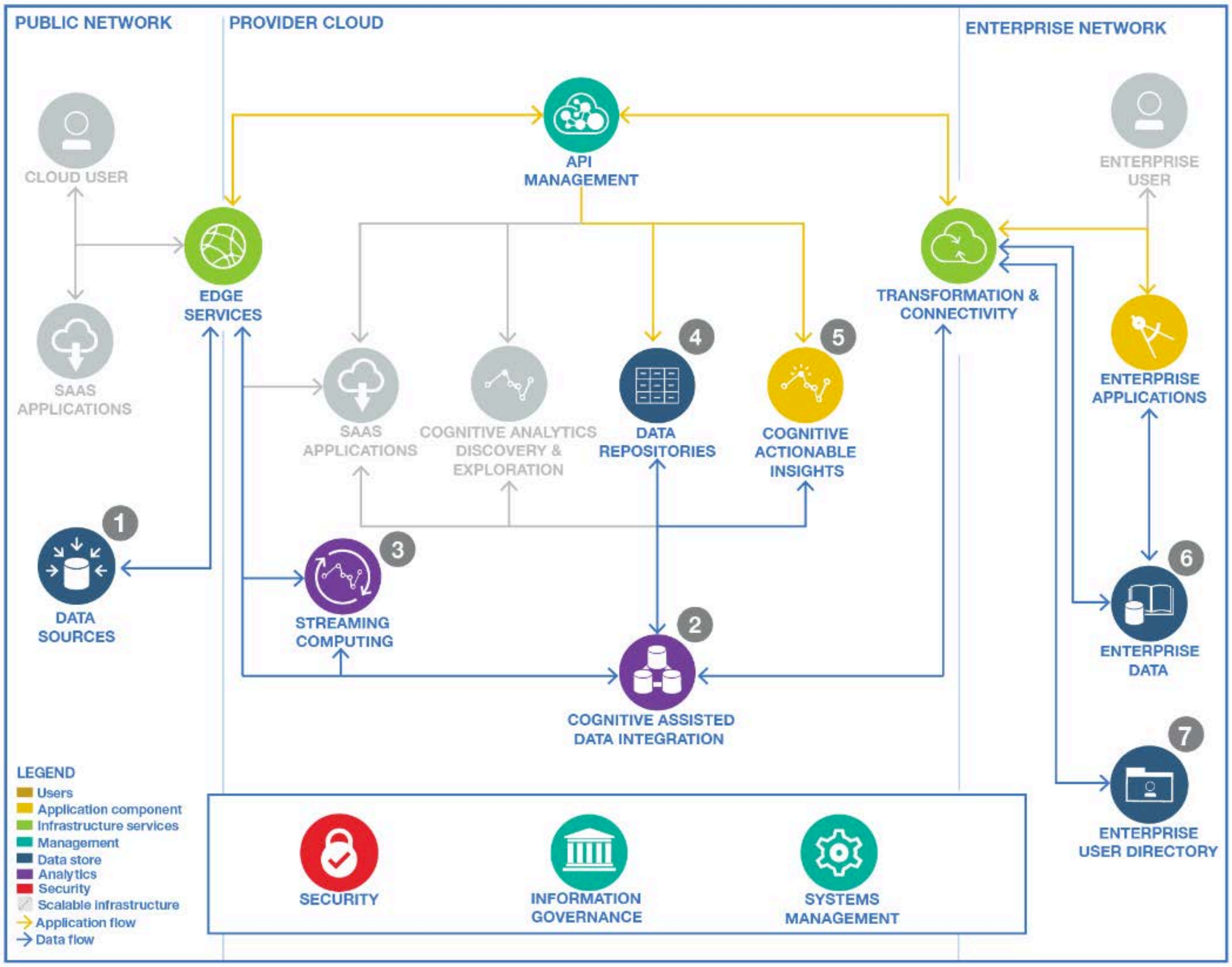


Figure 4: **Solution Architecture for Cyber Threat Intelligence**

Runtime Flow

1. Data is collected both from internal sources such as network probes, DNS, NetFlow, Active Directory (AD) logs, and network logs and from external sources such as blacklist/whitelist providers.
2. Most structured sources of data are sent first to the Security Information and Event Manager (SIEM), which acts as an integration layer and converts all incoming data into a single format.
3. Streaming Computing picks up both the streaming flow data (such as DNS and NetFlow) and the processed data from the SIEM system. It computes simple analytics (such as traffic in/out per server and number of requests made/failed to a DNS domain), which are later used in developing the machine learning models.
4. All raw data and output from Streams is sent to a data repository. Machine learning models are run against longer data sets to detect advanced persistent threats. Additional models are deployed in the data repository.

5. Machine learning models that have been developed in advanced analytics are deployed in 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 Active Directory (AD) logs are used for lookup to enrich and pinpoint user activity. An intelligence analytics tool is used to visualize the data from the data repository by the security analysts.
7. User lookup information is ingested from the enterprise Active Directory to establish exactly which user was involved in a particular traffic flow.

Functional Requirements

- The solution must support information coming in different formats from multiple data sources.
- The solution must be able to process real-time flow data such as NetFlow, DNS, and IP Flow.
- The vendor should have pre-defined parses to parse network flow traffic for the data sources mentioned above.
- The solution must be able to provide real-time threat indicators without relying on signature evaluation of malware.
- The solution must have analytics capabilities that can help users identify, correlate, and dynamically exploit emerging trends in data sets and data flows.

Summary

Hybrid cloud is one of the most common deployment models for big data and analytics. It is important for the business to have an overall strategy and operational roadmap in order to leverage hybrid cloud to enable efficiency, innovation, and growth. The level of sophistication depends on the actual business use cases and the IT environments including both existing and desired. Key deployment considerations should include data and analytics, data movement and replication, data preparation and integration, data sovereignty and compliance, data governance and security, HA and DR, network configuration and latency, workload and portability, scalability, and resource orchestration. A well-designed hybrid cloud solution/strategy will allow the enterprise to extend the on-premises investment to private and public cloud(s) and benefit from the capabilities each environment can best offer.

Works Cited

[1] Cloud Standards Customer Council 2017, *Cloud Customer Architecture for Big Data & Analytics.* http://www.cloud-council.org/deliverables/cloud-customer-architecture-for-big-data-and-analytics.htm

[2] ISO 17788 Cloud Computing Overview and Vocabulary. http://standards.iso.org/ittf/PubliclyAvailableStandards/c060544_ISO_IEC_17788_2014.zip

[3] Cloud Standards Customer Council 2017, *Data Residency Challenges.* http://www.cloud-council.org/deliverables/data-residency-challenges.htm

[4] Cloud Standards Customer Council 2016, *Practical Guide to Hybrid Cloud Computing.* http://www.cloud-council.org/deliverables/practical-guide-to-hybrid-cloud-computing.htm

[5] Cloud Standards Customer Council 2015, *Security for Cloud Computing: 10 Steps to Ensure Success* http://www.cloud-council.org/deliverables/security-for-cloud-computing-10-steps-to-ensure-success.htm