Quinnox
accelerate success

Data Warehouse and BI: On-Premise or On-Cloud

Quinnox Inc.
marketing@quinnox.com

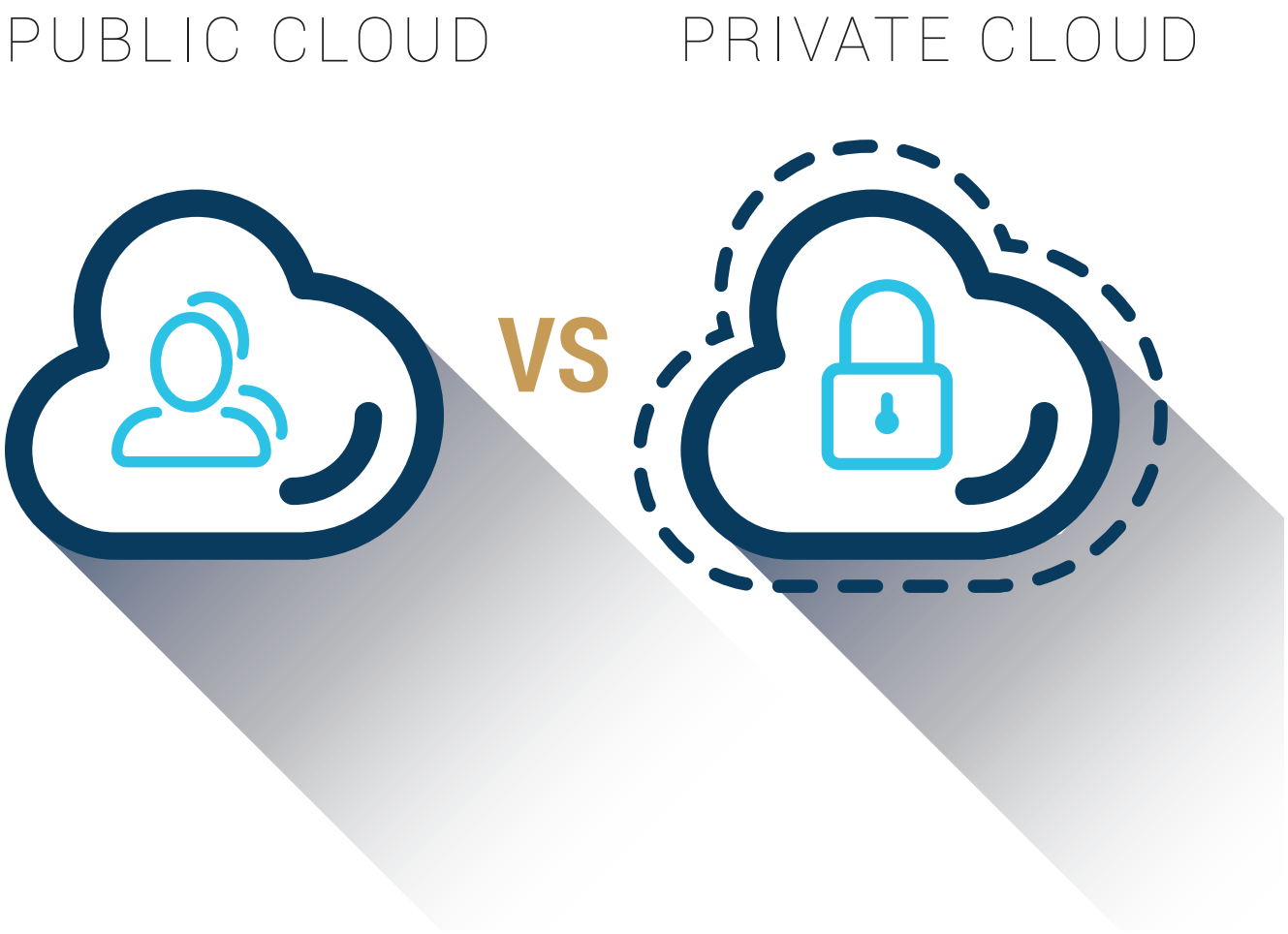
Introduction

Cloud computing is said to be the future of business technology. Companies like Amazon, Google and Microsoft have invested heavily in cloud computing, which has provided numerous opportunities for new business solutions. The benefits include immediate deployment, ease of scalability, high availability and low cost.

Currently, most of the cloud computing technology has been effectively used for deploying Web applications on the cloud. Traditionally, such applications support concurrent workloads with low latency response on transactions that are typically low in data volume and complexity.

Data warehouse and Business Intelligence workloads are read-intensive, so lots of data with high complexity (complex joins) is the norm. Moving from Business Intelligence to analytics, workloads are not only read-intensive, but also processing-intensive.

This raises an important question that needs careful analysis… Is cloud computing (both public and private cloud) a good fit for data warehouse and BI environments?

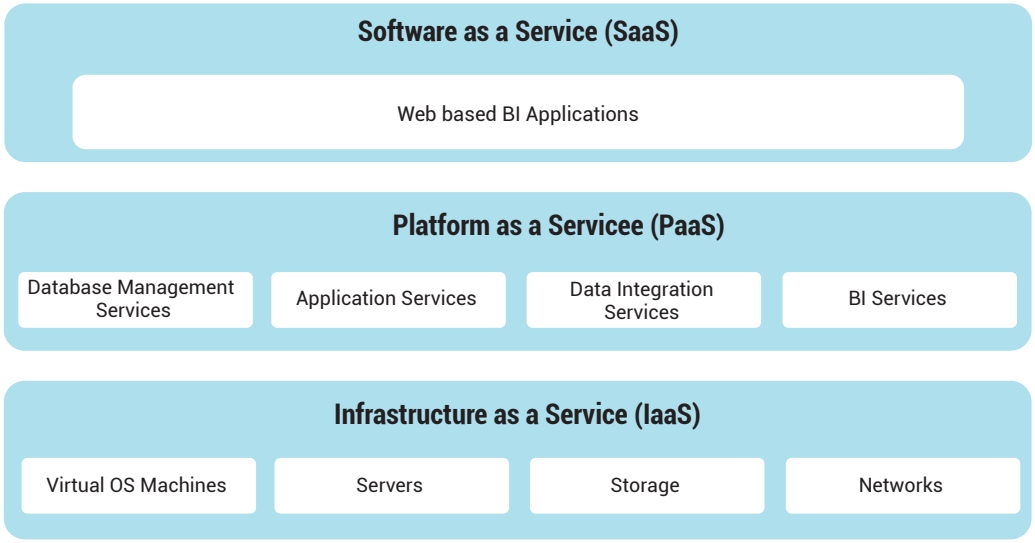


Cloud Computing Options

Cloud computing offers two options: public cloud and private cloud. Public clouds are owned and maintained by third-party cloud service providers. Public cloud benefits include cost reduction from shared resources as it encompass hardware, software, labor and maintenance across multiple organizations. However, concerns like data privacy, data workloads and security regulations prevent organizations from using a public cloud. Public clouds also provide another challenge related to data movement and data management between organization's internal systems and the public cloud.

Private clouds offer solution to these challenges. Private clouds are those that are built exclusively for an individual organization. While a private cloud solution addresses privacy, security and some workloads challenges, organizations do not benefit from sharing the resources model of public cloud, which can result in higher costs. Also when it comes to scalability, since public clouds are based on pay-for-use model, which helps in easy scaling of the environment, private clouds require organizations to procure hardware resources that result in additional cost.

Cloud Computing Service



Cloud computing services are generally categorized as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). All of these fall under the pay-for-use model. PaaS provides implementation platform on IaaS to develop SaaS web-based applications without the need to buy and maintain the underlying hardware, software and data center.

Why Use Data warehouse and BI in the Cloud

Low Costs

- There are number of factors that help to control cost when going for data warehousing in the cloud
- Cloud service providers purchase large number of hardware and software resources, so economy of scale helps to reduce per-unit costs
- There's an elastic environment model of the cloud that is usage-based and results in cost saving from reduced hardware resources when not needed. For example, there can be a release of development resources and test environments when not needed
- There are no costs related to hardware upgrades due to aging of physical components of the server.
- The cloud service provider will change the components at their own cost, and the entire process will be transparent due to virtual machines running on top of that server
- There are cost savings due to no requirement of a data center, server, storage or network

Performance and Scalability

Data warehouse and BI workloads are major concerns that require additional efforts by database administrators in order to effectively manage workloads and performance tune the environment. With the cloud, hardware resources are available on-demand with a pay-for-use model. When required for critical projects, these resources can be put to use without the need to procure hardware, or perform setup and configurations. This approach helps avoid necessary delays and results in immediate improvements in performance. This pay-for-use model also makes it unnecessary to plan for scalability, as required in traditional data warehouse.

Elasticity and Business Justification

The ability of the cloud to provide an environment for temporary or on-demand use makes is easy to reallocate IT resources to other projects or environments, or to greater priorities.

In addition, companies are business driven, so IT department are always under pressure to provide justifications in order to get approval for long-term projects. Cost, of course, is a major challenge to justify. In such a scenario, the IT team would execute a proof-of-concept project for an on-premise or public cloud environment. This would effectively illustrate the cost-effectiveness and value of the project—including predicted results—potentially making approvals easier and faster.

Challenges of Using Data Warehouse and BI in a Public Cloud

Workload Performance

Data warehouse and BI databases are designed to work with hardware to handle heavy workloads. The public cloud is built using commodity hardware limited by network speed between nodes, leading to performance bottlenecks. Again, in public clouds, though the CPU, memory and storage are dedicated, the virtual machines generally share same network hardware and I/O channels.

Data Integration

Most of the data within an organization is generated internally in a transaction system, then loaded into the data warehouse that is located in their data center. If this data from the transaction systems is to be loaded in an on-cloud data warehouse, then data movement through ETL workloads is a challenge due to internet network bandwidth bottlenecks. Also, such high internet usage may cause a negative performance impact on other applications that are dependent on internet bandwidth.

Privacy Laws

With the public cloud, data can be stored anywhere, even outside the geographic borders of a country. In such a scenario, if there are privacy laws about where data should be stored, having data on the public cloud can be a challenge for cloud service providers to maintain the data within specific geographic.

Data Security

Industries like financial services have concerns over data security and privacy making it an additional compliance issue for data on the public cloud. This is due to the public access nature of cloud and the sharing of hardware and software resources across multiple organizations. Also, security lapses, such as the HeartBleed bug and U.S. government's digital surveillance initiatives add more concerns around data security on the public cloud.

Single-node or Non-relational databases

A cloud is a collection of many small sized nodes that are combined together to form clusters, which further combine to form a virtual machine on which platforms and applications are built. If databases cannot grow beyond a single node, they won't be able to scale in a cloud environment. A cloud deployment needs a share-nothing and massively parallel processing (MPP) database that matches the cloud architecture of distributed hardware.

Similarly, non-relational, or 'NoSQL' databases, have limited SQL support, and the ability to join multiple tables. Data warehouse and BI, on the other hand, have always involved queries with complex joins across multiple tables.

Benefits of Private Cloud over Public Cloud

The challenges of the public cloud – workload performance, privacy laws and data security – can be managed using a private cloud. Workload performance on a private cloud is much better than a public cloud, since hardware can be configured for heavy workloads (plus the environment is in control of the organization deploying the private cloud). Also, since most private clouds are within the data center network bandwidth, privacy laws and data security issues are addressed. While addressing the challenges of public cloud, private clouds are still able to provide cost, performance, scalability and elasticity benefits. The private cloud does provide the flexibility to address variable workloads by better using hardware resources with elastic facilities.

	Traditional Data warehouse/BI	Public Cloud Data warehouse/BI	Private Cloud Data warehouse/BI
Initial costs	High	Low	Medium to high, but depends on vendor appliance, or whether the full environment is procured
Incremental Costs	High	Low	Low, until hardware limit is reached
Environment Lead Time	High	Low	Low, provided all hardware is procured and virtual machines are configured
Scalability	Need to plan	Immediate	Need to plan
Workload Performance	Heavy workloads	Low workloads	Heavy Workloads
Data Integration	Easy	Challange	Easy
Privacy Laws	Comply	May or may not comply	Comply
Data Security	High	Low	High

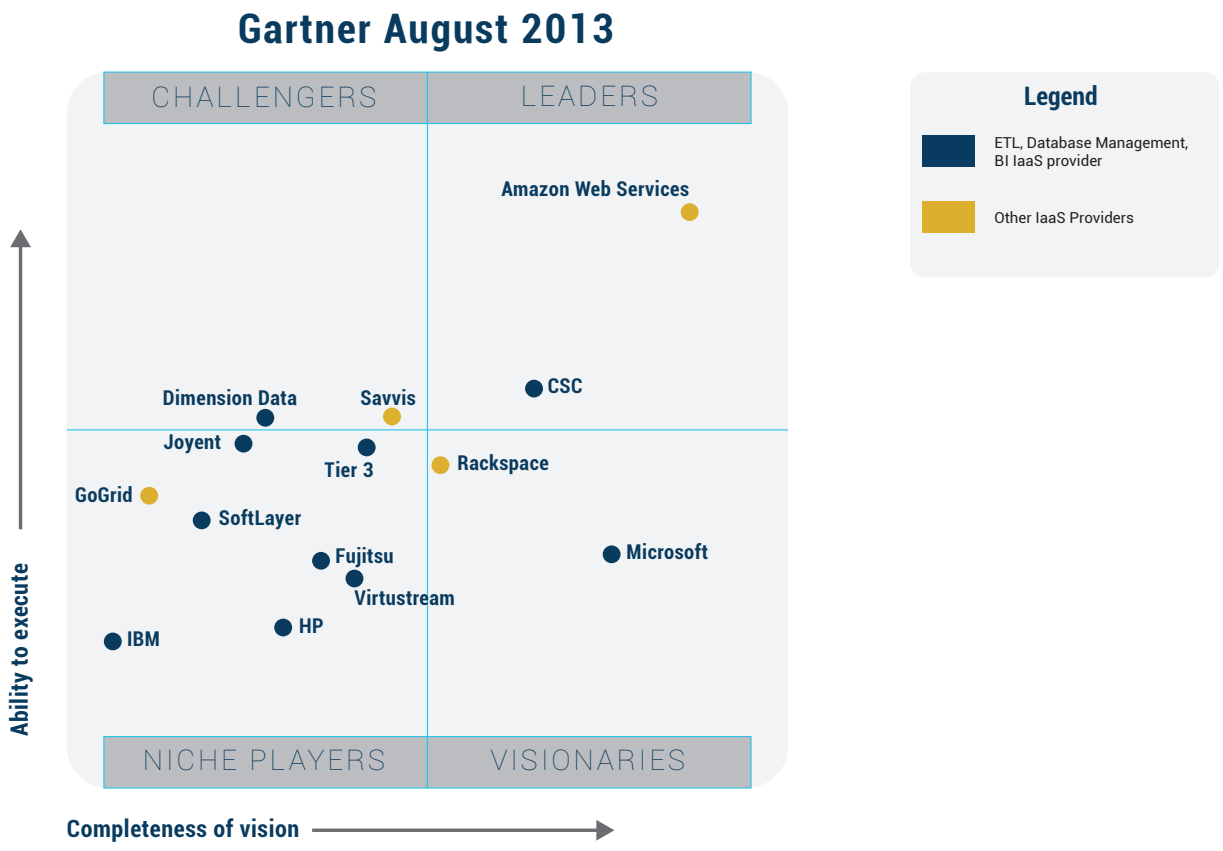
The challenge of data integration for public cloud data warehouse and BI environments can be effectively managed using data virtualization. Data virtualization offers a solution that avoids data movement between organization's internal systems and the cloud. This is done by combining disparate data sources into a single "virtual" data layer that provides unified access and integrated data services to consuming applications.

Data virtualization is much more than data integration; it provides advanced features like virtual data profiling, data quality improvements, advanced caching and more. The decision to go with data virtualization can be purely made by evaluating data consuming needs based on latency, hardware and network resource constraints, and committed SLAs.

Service Providers

Data warehouse / BI IaaS Providers

Amazon, Savvis, Rackspace and GoGrid are offering a pay-for-use model on hardware and Virtual OS machines on which companies can buy and deploy their own ETL, Database Management and BI software. There, however, may be some limitations in terms of what software the service providers may maintain and support. This needs to be evaluate based on the environmental requirements.



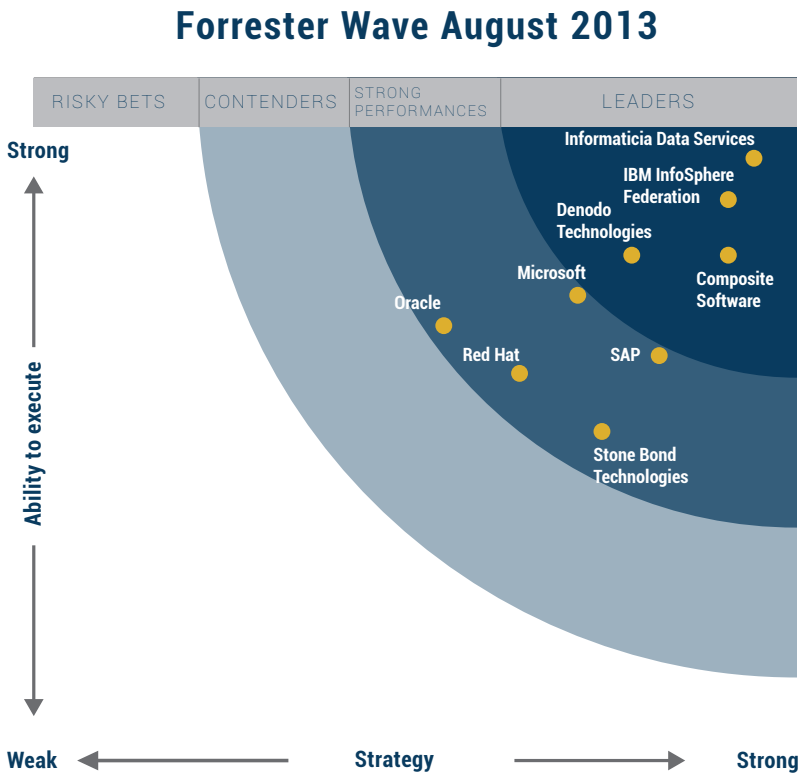
Data Warehouse / BI PaaS and SaaS Providers

Date warehouse and BI PaaS and SaaS provider can be broken down into three categories: date warehouse, business intelligence and data integration. Below is a sampling of some of the more established players in these three areas.

PaaS/SaaS Options	Service Providers
Data Warehouse	Amazon Redshift IBM DB2 on Amazon EC2 Microsoft SQL Server on Windows Azure MySQL in Cloud Oracle on Amazon RD5 SAP HAHA Cloud Platform Teradata Express on Amazon EC2 Vertica Analytic Database on Amazon EC2
Business Intelligence	ActuateOne BIRT iHub onDemand IBM Cognos Platform Jaspersoft Cloud Analytics on Amazon EC2, GoGrid and Microsoft MicroStrategy Cloud QlikMarket on Amazon EC2 SAP BusinessObjects BI OnDemand Tableau Online Tibco Spotfire Cloud
Data Integration	Pentaho Data Integration IBM InfoSphere DataStage on Amazon EC2 IBM Integration Bus Hypervisor (formerly WebSphere Informatica Powercenter Cloud SoftwareAG webMethods CloudStreams Talend Integration Platform

Data Virtualization Providers

According to Forrester, companies fall under one of four categories based on current offerings and overall strategy in the data virtualization space. Here's an illustration of some recent findings in this area.



Amazon's Redshift – on Cloud Data Warehouse

Amazon's Redshift provides several advantages, including cost savings over traditional data warehouse—around $1,000 versus $19,000-$25,000 per terabyte per year. This service also offers a cloud data warehouse solution that is:

- Built on massive parallel processing (MPP) data warehouse ParAccel
- Designed to scale from terabyte up to multiple petabyte size
- Optimized for columnar data storage and structured data, plus takes advantage of advanced compression techniques
- Easy to integrate with Amazon's DynamoDB NoSQL database and Amazon's Simple Storage
- Service (S3) for load/dump to/from Redshift tables
- Supportive of the AWS Data Pipeline to transfer of data all around AWS's cloud, using a 10 gigabit connection

Conclusion

On-cloud data warehousing certainly provides a cost-effective model in comparison to traditional data warehouses. The public cloud is maturing as a platform for supporting applications; however, in terms of data warehouse and BI workloads, public cloud is still immature.

A private cloud solution provides many benefits of a public cloud, in terms of scalability and elasticity. It also offers the benefits of traditional data warehousing, in terms of workload performance, data integration, privacy and data security. However, a private cloud solution does not offer the same cost benefits as the public cloud, since it is not based on a pay-for-use model. What private cloud solutions do offer is the ability to help organizations fine tune cloud best practices and processes, which can be a valuable “get-ready” phase as the public cloud continues to mature.

References

http://intelligentbusiness.biz/wordpress/?p=281
http://virtualization.sys-con.com/node/1833397