02 Storage Technologies for AI, Big Data and the Cloud

www.huawei.com

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

https://t.me/learningnets

Foreword

- This module mainly introduces:
  - Development trends of new ICT architectures.
  - Concept of the Cloud, the key technologies used and the application of storage in the Cloud.
  - Concept of Big Data, the key technologies used and the application of storage in Big Data.
  - Converged technologies and applications between the Cloud, Big Data and AI.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

https://t.me/learningnets

Objectives

- Upon completion of this module, you will be able to:
  - Understand the development trends of ICT.
  - Understand what is the Cloud, Big Data and AI.
  - Learn about storage technologies and its application in the Cloud.
  - Learn about storage technologies and its application in AI and Big Data.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

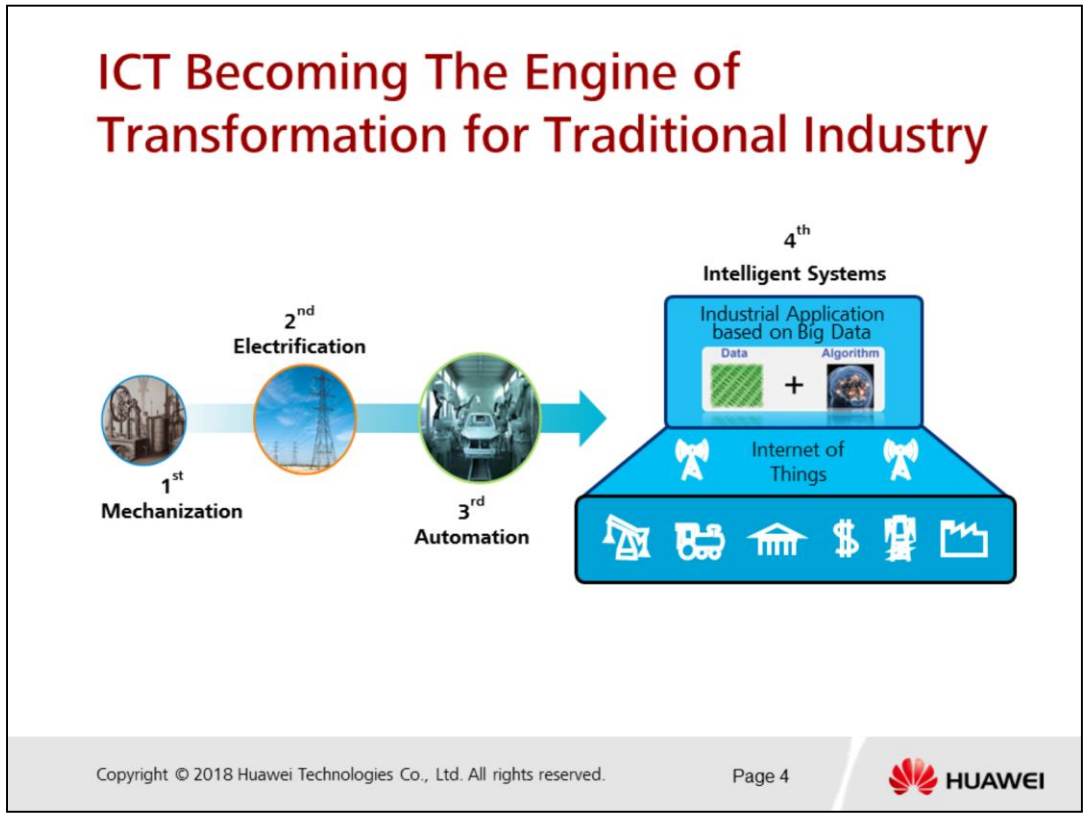
https://t.me/learningnets

Contents

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

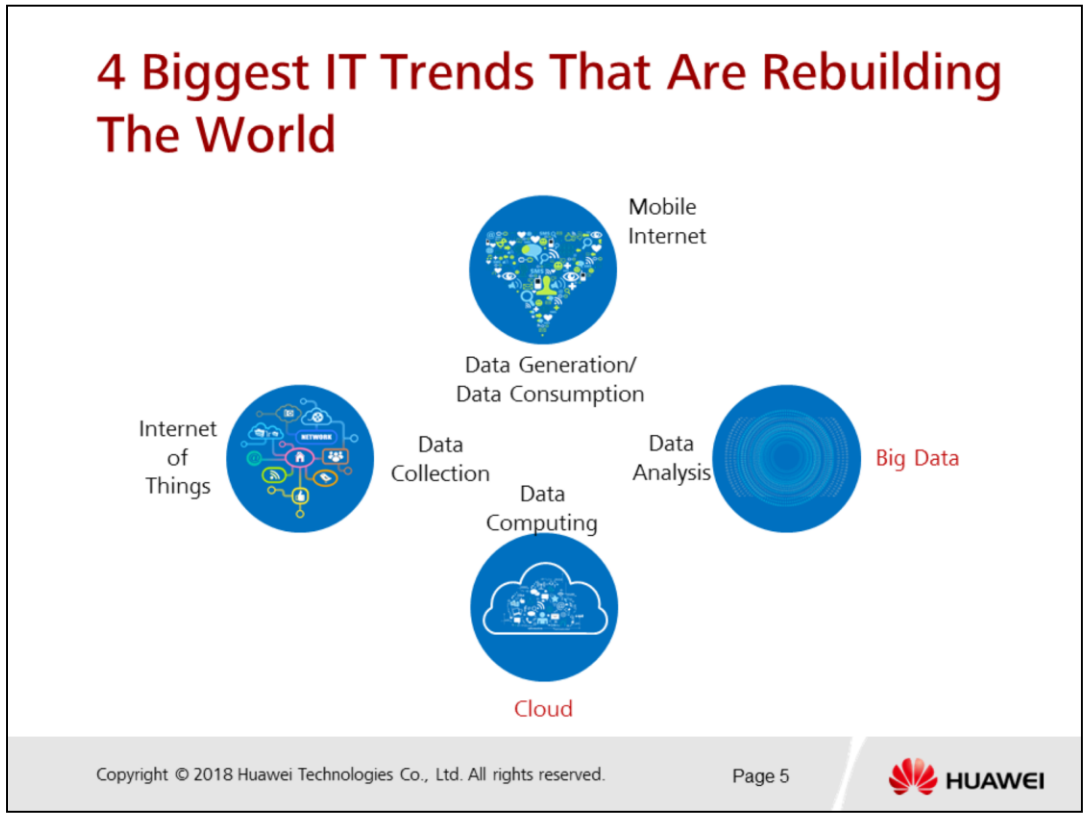
HUAWEI

https://t.me/learningnets



- All types of industries are facing the challenges of new IT technologies, and the new trends in IT are changing the world and rebuilding the rules that forms the basis of the digital world. As new technologies emerge, new industries are formed and there industries that are being replaced by these technologies, we can see the effect of technology on the world by looking at the timeline of the Industrial Revolution.
- The First Industrial Revolution used water and steam power to mechanize production. The Second Industrial Revolution used electric power to create mass production. The Third Industrial Revolution used electronics and information technology to automate production. Now a Fourth Industrial Revolution is building on the Third, the digital revolution that has been occurring since the middle of the last century. It is characterized by a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres to create intelligent systems. Some of the intelligent systems such as industrial application based on Big Data, and Internet of Things are key elements that will bring forward the Fourth Industrial Revolution.

https://t.me/learningnets

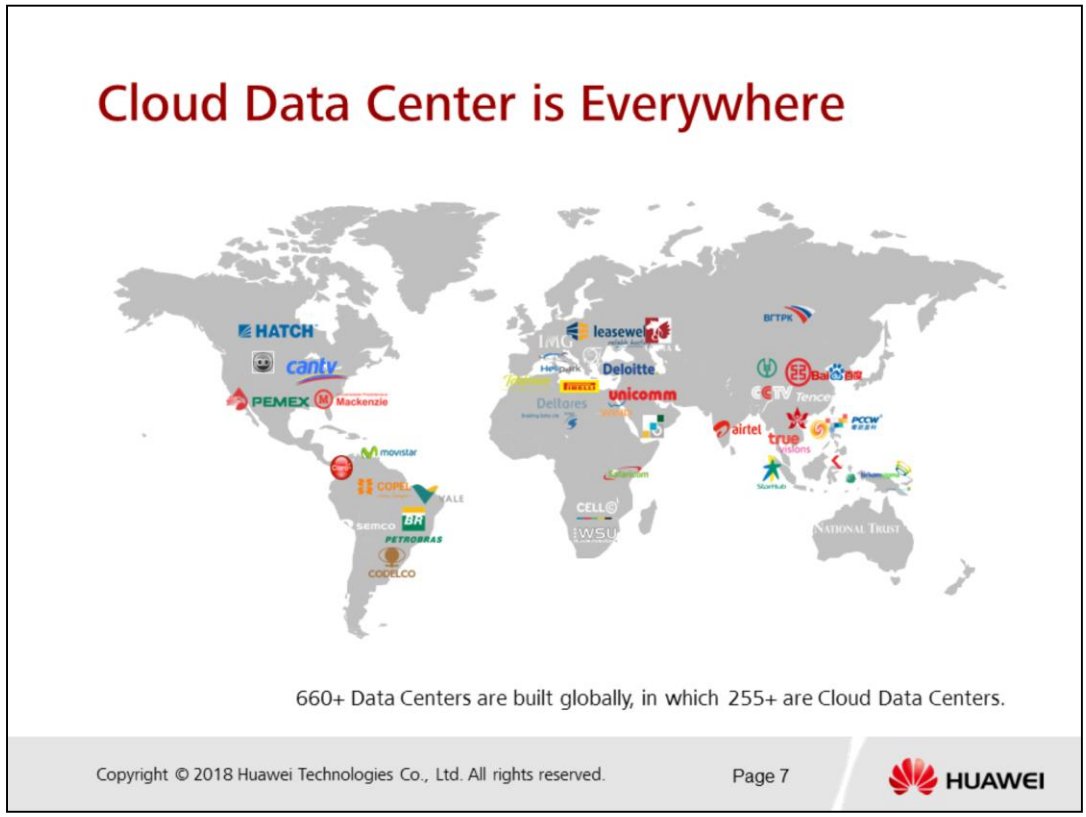


- Mobile Internet: 7 billion mobile users, which is near to the total population on Earth. 78% CAGR of data growth and the total number of smart phones in the world (1.82 billion) has exceed the number of PC (1.78 billion). The compound annual growth rate (CAGR) is the mean annual growth rate of an investment over a specified period of time longer than one year.
- Internet of Things: Social Business and Social Media changed the people's way of life. 86% of enterprises are developing their business in social media sites.

https://t.me/learningnets

- Big Data: Data is one of the most important asset of businesses, and enterprises will be competing between each other in terms of data in the future. McKinsey (a worldwide consulting firm) mentioned that enterprises that are unable to fully utilize the capabilities of Big Data will be phased out in the future. Big data analysis has given us a lot of promises, but in the near term, there are just too many big data solutions looking for solvable problems. In the long run, the potential of big data will outpace the optimization of e-commerce by embracing all vertical sectors, including the financial sector, manufacturing, transport and power sectors etc. However, all these vertical sectors requires Industrial Internet(also known as Internet of Things) to interconnect large amount of sensors that could provide huge amount of data that can be used for improving product designs, and accurate prediction of faults etc. GE (General Electric) and IBM are the pioneers and leaders of this field currently, but we are still at the very initial stages of Big Data revolution. In a few years from now, Industrial Internet will develop in a greater pace and Big Data will also grow bigger, making the demands for Big Data solutions unstoppable.
- Cloud: The Cloud has become the next generation of IT infrastructure, 56% of SMEs will purchase 4+ of cloud services in the next 3 years. Up to 2016, 75% of the new IT investment is on the Cloud or Hybrid Cloud. 70% of the CIOs also deploys a "Cloud Prioritization" strategy in 2016. 80% of the new IT decisions will have the business representatives involved, while 53% of the IT decisions will be led by the business representatives.

https://t.me/learningnets



- Market trends has a huge change from prioritizing internal resources to external resources. Back in the days, enterprises are focused on building own data centers and server rooms, but slowly the market trend changed to using external services and resources, which promoted the growth of data centers worldwide. Data centers as a service are very common nowadays, and businesses no matter big or small can purchase resources from data center service providers worldwide.

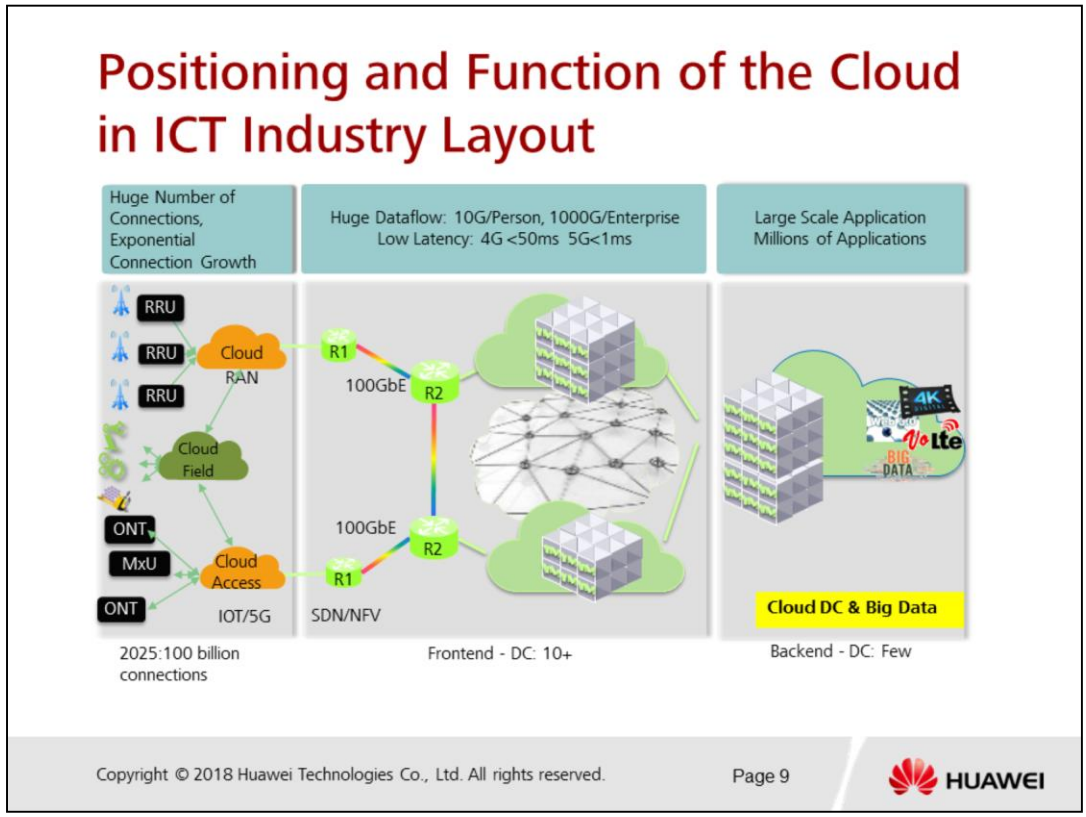
https://t.me/learningnets

Contents

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

https://t.me/learningnets

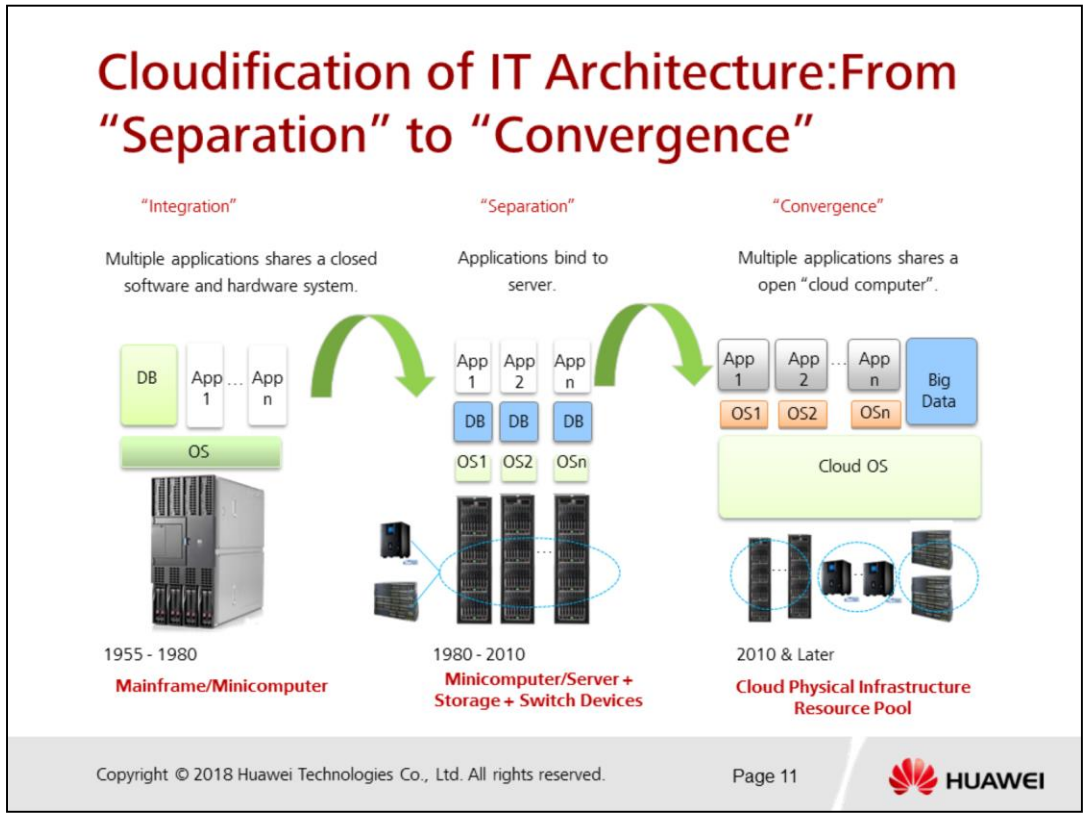


- The whole ICT industry layout has been set and just waiting for the maturity of the big modules and interconnection of those modules. Technologies within the modules such as 5G, NFV/SDN, Cloud Data Center, Big Data are growing in a very quick pace.
- Cloud Data Centers faces 3 big challenges: huge increase in connection flow, huge increase in data processing, and huge increase in application services.
- Internet of Things (IoT): It is the interconnection via the Internet of computing devices embedded in everyday objects, which enables them to send and receive data. The internet of things (IoT) is also a computing concept that describes the idea of everyday physical objects being connected to the internet and being able to identify themselves to other devices, and exchange data between them.
- SDN (Software Defined Network), is an emerging network architecture that decouples the network control and forwarding functions, which enables the network control to become directly programmable and the underlying infrastructure to be abstracted for applications and network services. In a software-defined network, a network administrator can shape traffic from a centralized control console without having to touch individual switches, and can deliver services to wherever they are needed in the network, without regard to what specific devices a server or other hardware components are connected to. The key technologies for SDN implementation are functional separation, network virtualization and automation through programmability.

https://t.me/learningnets

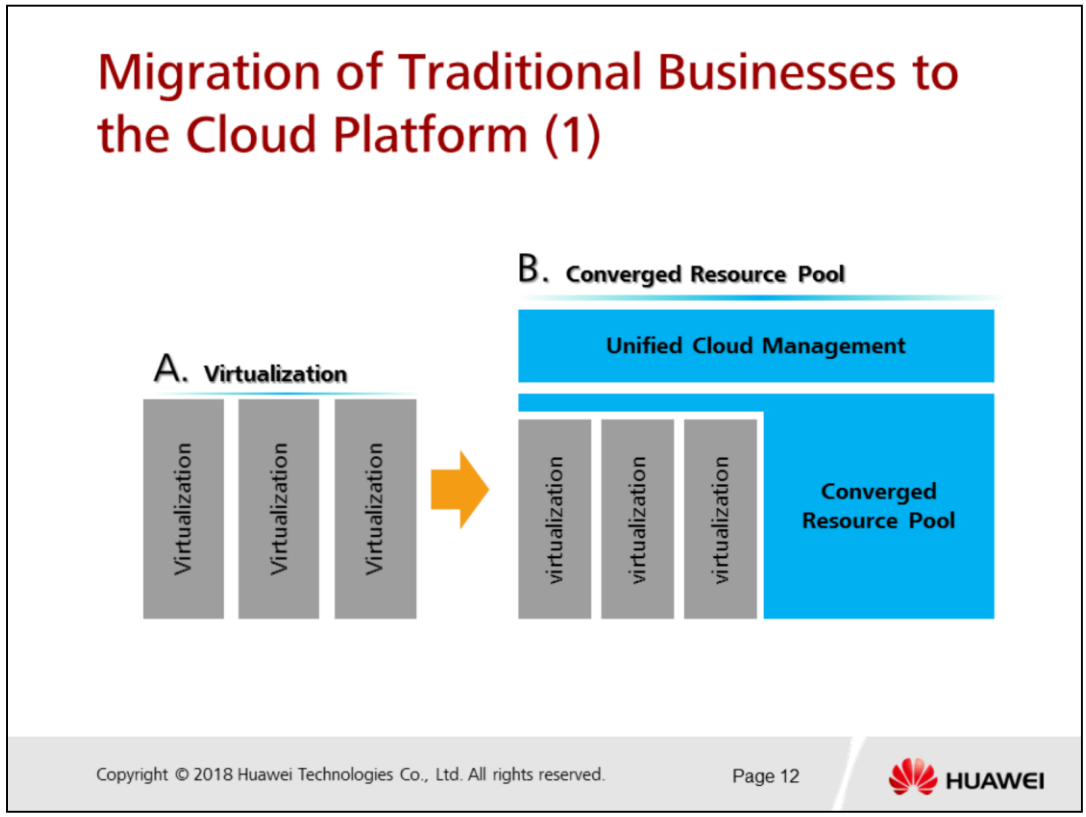
- NFV (Network Function Virtualization), is a network architecture concept that uses the technologies of IT virtualization to virtualize entire classes of network node functions into building blocks that may connect, or chain together, to create communication services. NFV decouples the network functions, such as network address translation (NAT), firewalling, intrusion detection, domain name service (DNS) and caching etc. from proprietary hardware appliances so they can run in software.
- Cloud Radio Access Network (also known as Cloud RAN, C-RAN or Centralized RAN), is a new cellular architecture for future mobile networks. It is a centralized, Cloud-based architecture for radio access networks that supports 2G, 3G, 4G and future wireless communication standards.

https://t.me/learningnets



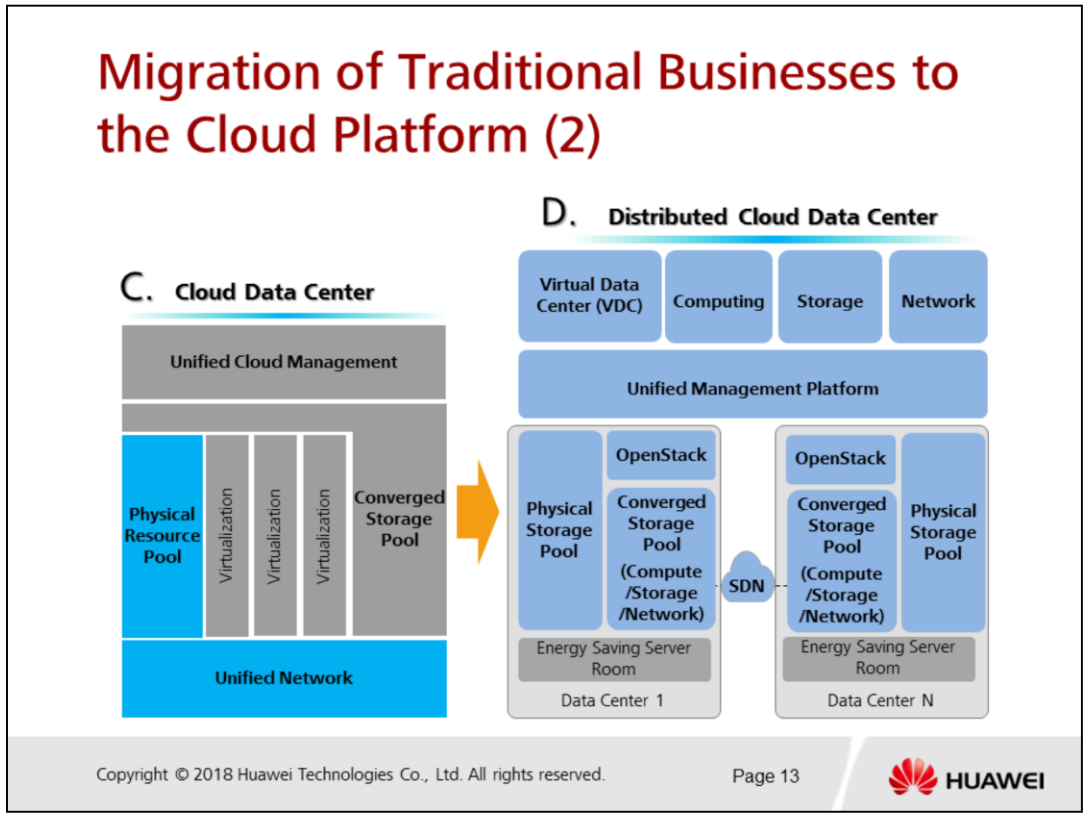
- Convergence of physical infrastructure: Heterogeneous computing, storage and network devices in the data center are unified and converged via cloud operating system with automated resource management.
- The driving force of value: Lower TCO(Total cost of ownership), higher efficiency in service deployment, more agile response towards client's requirements.

https://t.me/learningnets



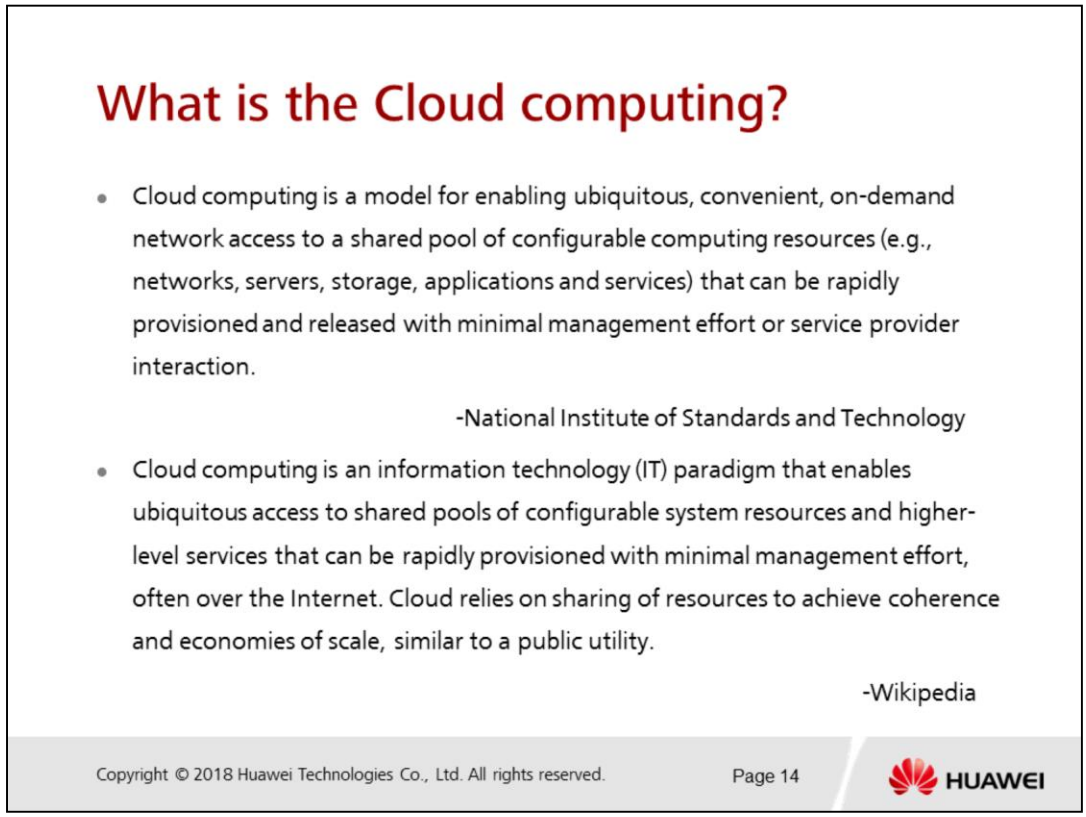
- Infrastructure:
  - Interconnection and evolution of existing networks.
  - Security isolation and security protection.
  - Building a differentiated resource pool that is able to satisfy different business needs.
- Management:
  - Management of existing resource pools.
  - Unified management of virtualized and physical resources.
  - Set resource pool management standard practices, standardized use of resources.
- Migration of services:
  - Smooth migration of old physical machines and data.
  - Smooth migration of virtual machines and data.

https://t.me/learningnets



- Virtualization is the process of running a virtual instance of a computer system in a layer abstracted from the actual hardware. Most commonly, it refers to running multiple operating systems on a computer system simultaneously. To the applications running on top of the virtualized machine, it can appear as if they are on their own dedicated machine, where the operating system, libraries, and other programs are unique to the guest virtualized system and unconnected to the host operating system which sits below it.
- SDN encompasses multiple kinds of network technologies designed to make the network more flexible and agile to support the virtualized server and storage infrastructure of the modern data center. Software-defined networking can be defined as an approach to designing, building, and managing networks that separates the network's control or SDN network policy and forwarding planes thus enabling the network control to become directly programmable and the underlying infrastructure to be abstracted for applications and network services for applications as SDN, Cloud or mobile networks.
- OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, which is all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface.

https://t.me/learningnets



- Cloud refers to the addition, usage and delivery model of related internet-based services that typically involves the provision of dynamically scalable and often virtualized resources over the Internet. Cloud is a metaphor for the network and the Internet. In the previous diagrams, we often used a cloud to represent communication network, and even use it to represent the abstraction of the Internet and the underlying physical infrastructure. The narrow definition of the cloud refers to the delivery and usage model of IT infrastructure, in which resources that you require can be gained through the network based on demand in a easily scalable method. The wider definition of the cloud refers to the delivery and usage model of services, where the services you need can be obtained on demand and in a easily scalable way through the network. These services can be IT, software or internet related services or any other services. This in a sense also means that computing power can also be circulated over the internet as a form of product or commodity.
- Cloud resources are dynamically scalable, virtualized and provided over the Internet. End users do not need to know the details of the infrastructure in the cloud, and they do not have the expertise in cloud technologies, and they also do not need to directly control them, in fact they should just focus on what resources they really need and how they will be able to get the service over the network.
- The cloud provides a layer of abstraction between the end users and the physical infrastructure that is running the cloud. This makes it easier for end users to get the resources and services they need without the hassle of needing to learn or have the expertise in the multiple technologies required to run a cloud.

https://t.me/learningnets

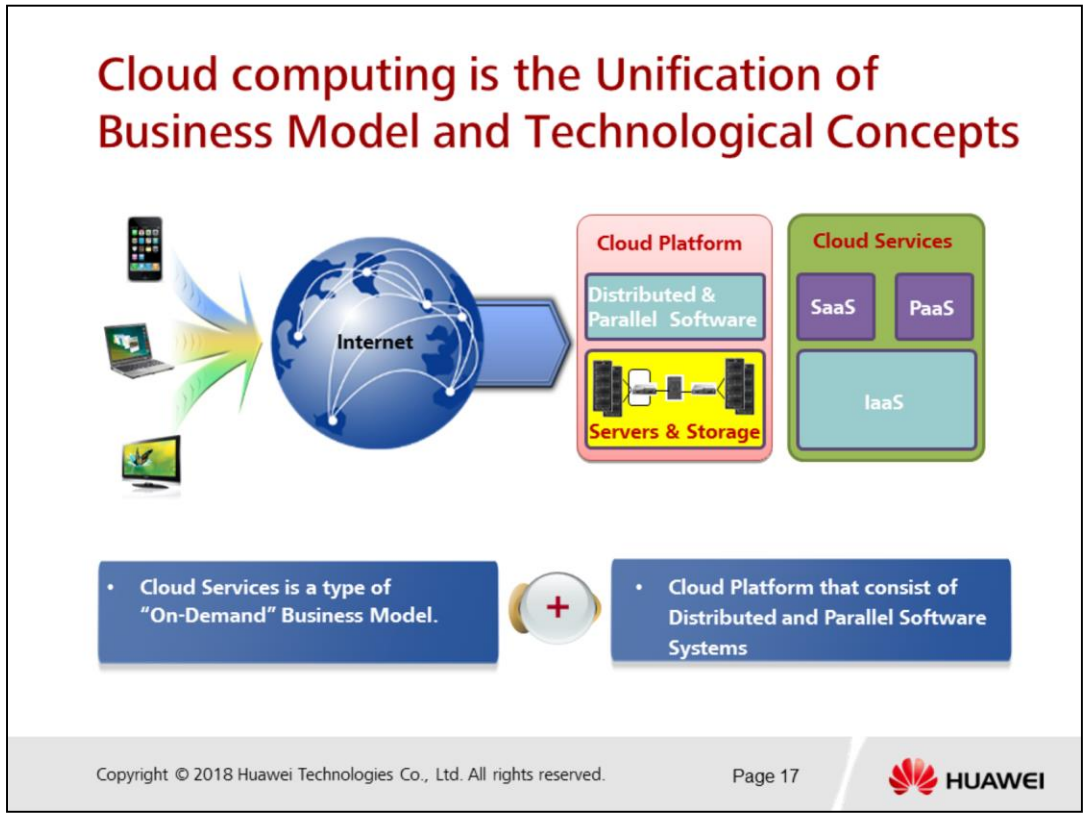
- Key Characteristics of the Cloud:
    - On-demand Self-service: Consumers can deploy the processing power and resources such as servers and network storages based on demand without the need of human interaction with the service providers of each resources.
    - Ubiquitous Network Access: Various capabilities or functions can be performed over the Internet through standard means of access on different types of client connection devices such as mobile phones, laptops, PDA etc. Your resources and services on the cloud easily accessed and managed over the internet through standard means without any complicated network connection setup.
    - Location Independent Resource Pooling: The service provider's resources are centralized, making it easier to provide services to customers in a multi tenant rental basis, and at the same time it is able to dynamically allocate differentiated physical and virtual resources based on customer requirements.
    - Rapid Elastic: Can rapidly and elastically provide resources and scale up or down based on demand. From the viewpoint of the customer, it seems that the resources that can be rented is unlimited and they are able to change the amount of the resources at any given time.
    - Pay per user: The billing for the resources are on pay per use basis or advert basis, which promotes optimum resource utilization. For example, storage, bandwidth, computing resource consumption can be billed monthly based on the user's actual usage. In an enterprise cloud, resource consumption can be billed by departments based on their usage.

https://t.me/learningnets

Cloud deployment modes: Private Cloud, Public Cloud, Hybrid Cloud.

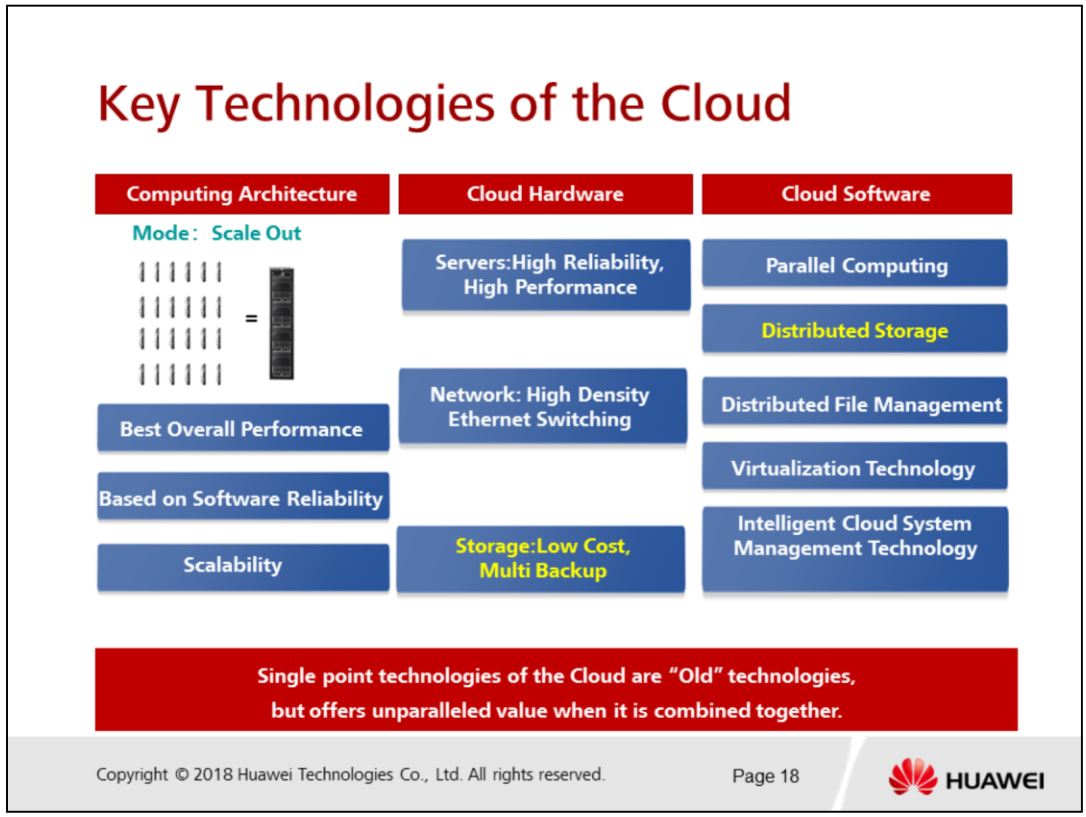
- Private Cloud: It is generally is used within an organization and at the same time managed by the organization itself. Huawei data center is an example of private cloud as Huawei is the cloud provider and at the same time is also the cloud consumer.
- Public Cloud: Just like a shared network switch, the telecom service providers operates and manages this switch but the actual users are the peoples of the public. Generally, public cloud provides cloud services to more than 1 organizations publically which means that anyone can opt for purchasing cloud services from a public cloud.
- Hybrid Cloud: This refers to deployed infrastructure of the cloud that consist of 2 or more types of Cloud, but is presented to the users as a single cloud. In the normal operation of an enterprise, critical and important corporate data (e.g. financial data) is stored within the private cloud while the unimportant data is stored on the public cloud, and these 2 clouds joins together to form a whole complete cloud system which is the hybrid cloud. For an example, lets consider an ecommerce website that has stable business volumes during normal seasons purchased the equipment to build a private cloud to handle the business services. However, during the promotional season or holiday seasons, the business volume and transactions spikes higher than usual. In order to cope with these spikes in business transactions, they paid the public cloud providers to rent some cloud servers to balance the service workload during these periods. Although some of the resources are from the public cloud, but they are still able to uniformly manage and dynamically allocate these resources, which makes this setup a form of hybrid cloud.

https://t.me/learningnets



- "On-Demand" business model is a business model where services are obtained based on customer's demand. The application and data required by the customer resides in the cloud, and the users use clients to access the required applications based on their demand. Cloud service providers provide the corresponding services to the customers based on their demand and collect corresponding payments for the services offered to the customers.
- From a business perspective, cloud services can be categorized as the following types:
  - IaaS (Infrastructure as a Service): IaaS provides the ICT infrastructure and resources that ensures normal operation of the services.
  - PaaS (Platform as a Service): PaaS provides a development platform for developers. In a traditional development environment, the developer needs to focus on the operating system and hardware while writing the codes for an application. However, with PaaS, the developer can focus on writing the codes for better applications while the infrastructure and platform is managed by PaaS vendors.
  - SaaS (Software as a Service): This is the earliest service model in the cloud. Users just require a simple device to connect to the application and operating system provided by the SaaS vendor. All the complex system maintenance tasks such as software and licensing upgrades are managed by the SaaS vendor.

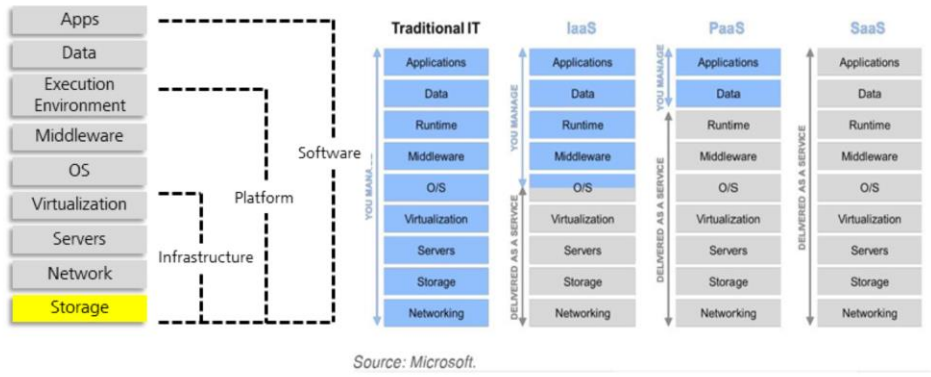
https://t.me/learningnets



- By summarizing multiple core technologies, it can be summed up in 3 main aspects which are: overall computing architecture, hosted hardware devices, and software system.
- The overall computing architecture requires high performance, high reliability and high scalability.
- Cloud hardware includes: High reliability and high performance computing servers that provides the computing resources. Low cost and data safe storage devices that provides data storage space. Supports data communication and exchange via high density switches in large layer 2 networks.
- Cloud software includes: parallel analysis computing technology used for Big Data. It also includes distributed storage technologies that provides dynamically scalable resource pool through converged storage resources, distributed file management used for data management and virtualization technologies for resource pooling of computing and storage. It also simplifies the operation and maintenance personnel's work through highly efficient and intelligent O&M system management technologies.

https://t.me/learningnets

Cloud Service Models



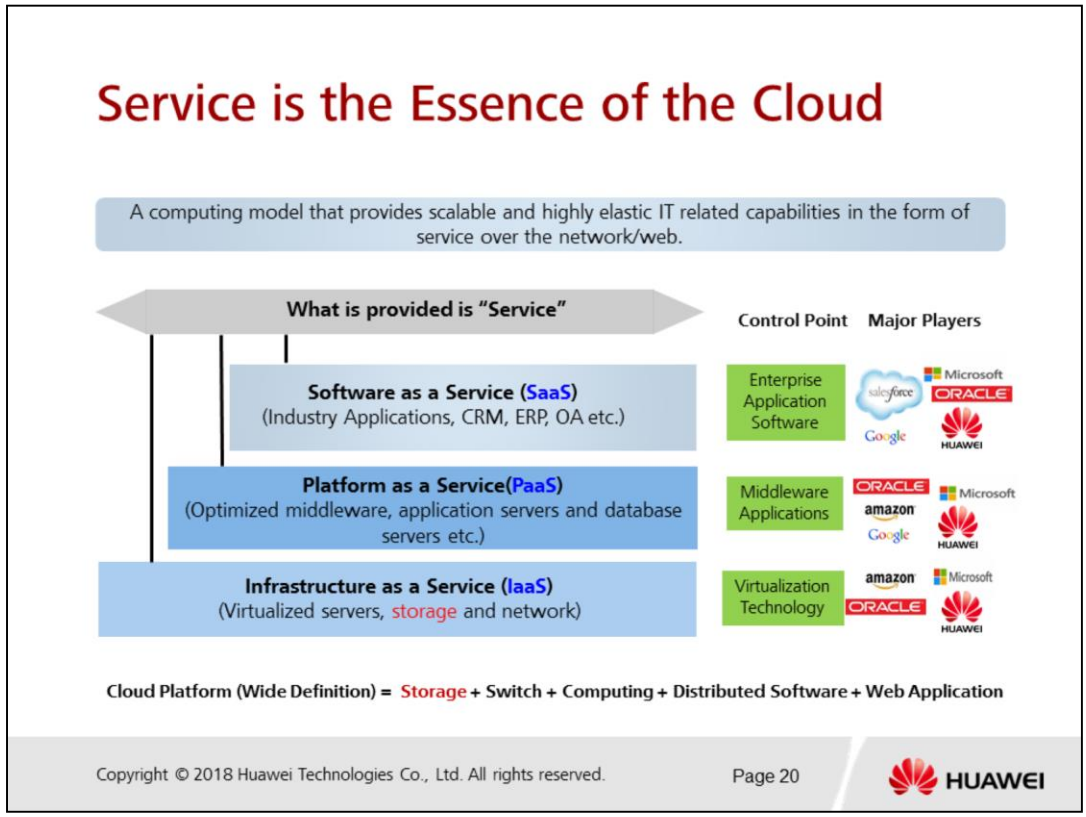
Source: Microsoft.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

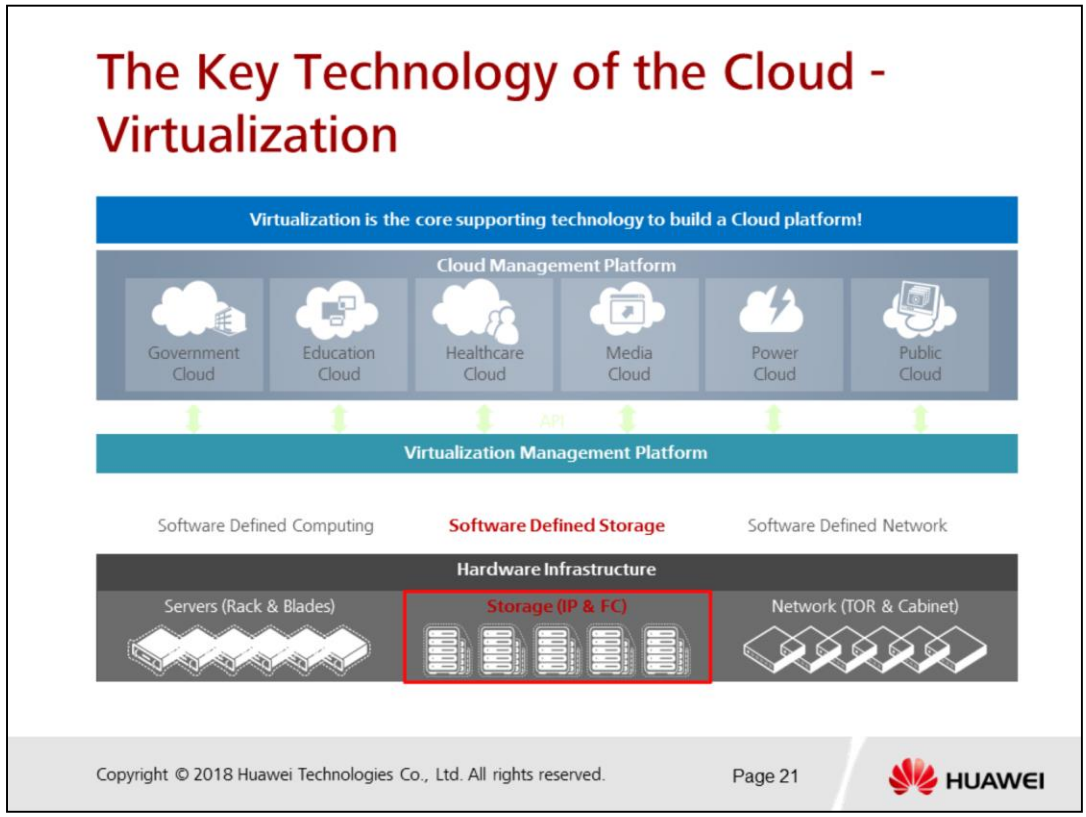
- The key characteristics of Cloud is:
  - On-demand Self-service
  - Ubiquitous Network Access
  - Location Independent Resource Pooling
  - Rapid Elastic
  - Pay Per Use

https://t.me/learningnets



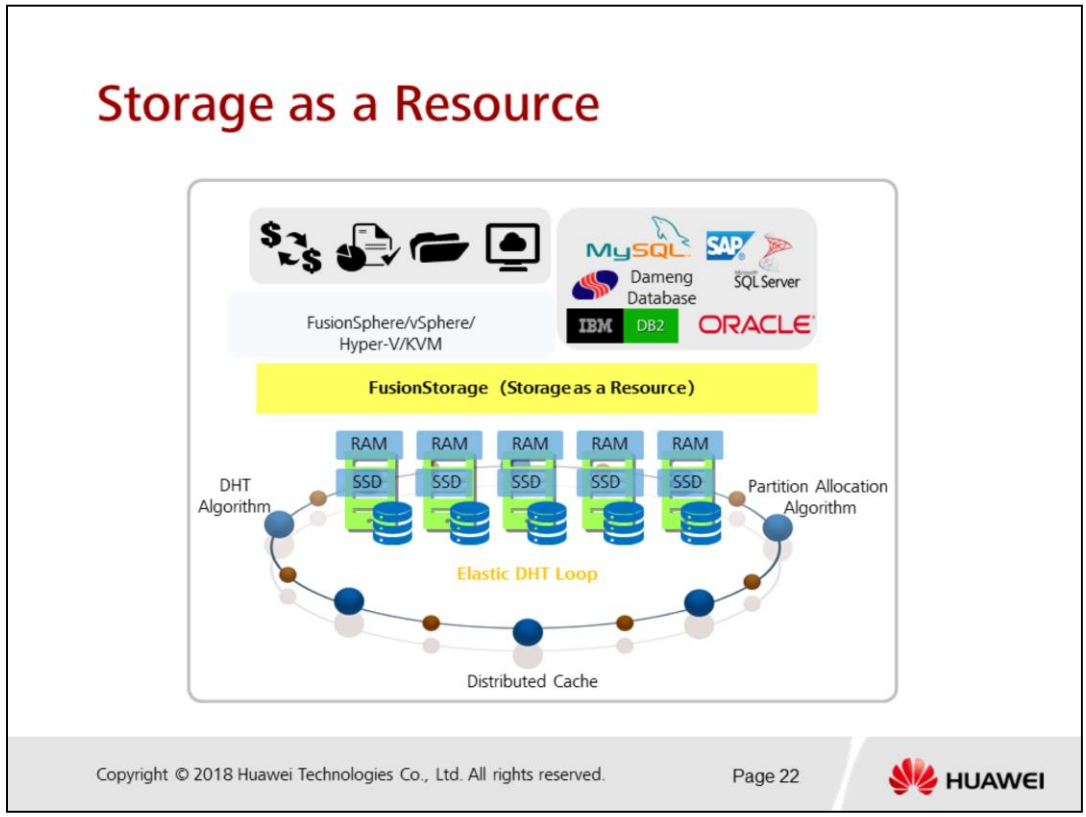
- The basic principle of a "Cloud Platform" is a distributed computing system, it provides On-Demand supercomputing and storage capabilities by distributing the computing load across a large number of computers. The terminal is also a part of the Cloud.
- The core part of the cloud platform is still the data center, but it is different in terms of technologies if compared to a traditional data center. Its grows in the direction that emphasizes standalone performance and clustering towards a "distributed, intelligent, and large capacity" data center.
- "Cloud" is just a metaphor, which refers to the networked distributed computing devices, and it also refers to the fact that data is hidden during the calculation process.

https://t.me/learningnets



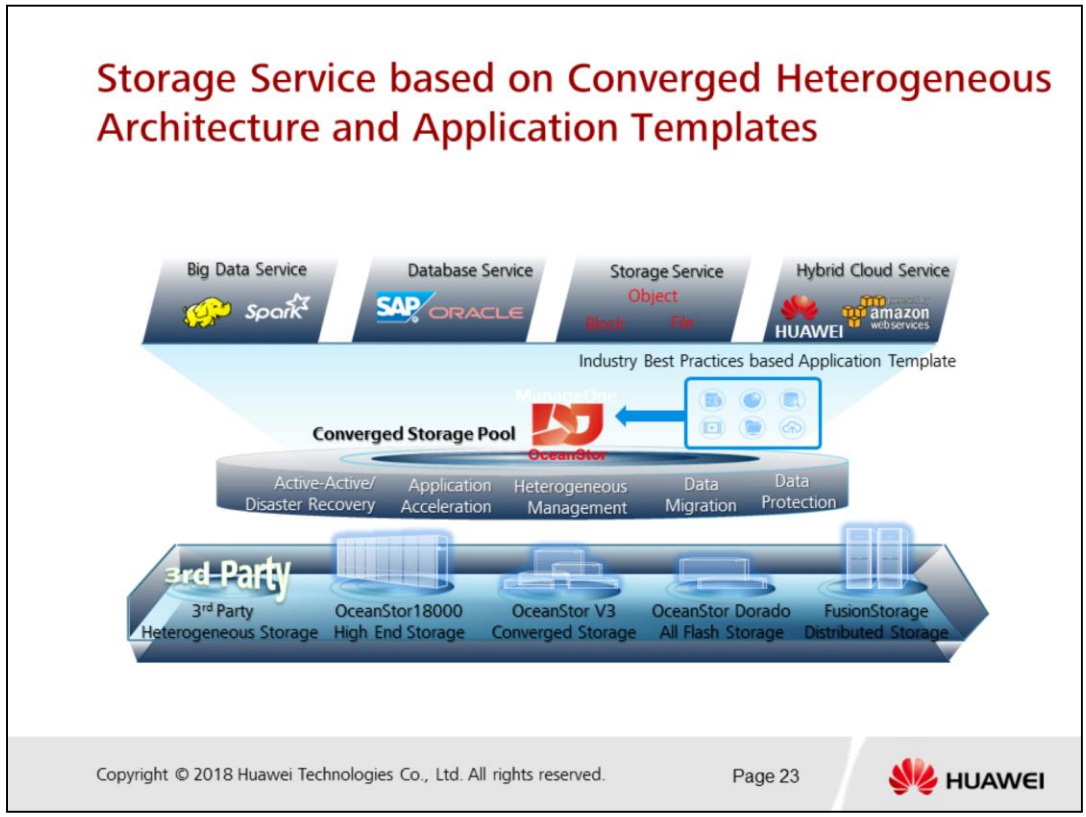
- Definition of Virtualization: Abstraction and isolation management technology towards physical resources in order to achieve the purpose of highest utilization rate of the physical resources.
- Types of Cloud deployment:
  - Private Cloud: Enterprise owned independent cloud infrastructure.
  - Public Cloud: Cloud infrastructure owned by cloud service providers that provides cloud service to the public or enterprises.
  - Hybrid Cloud: Combination of private and public cloud infrastructure through dedicated connection technologies to achieve data and application sharing.

https://t.me/learningnets



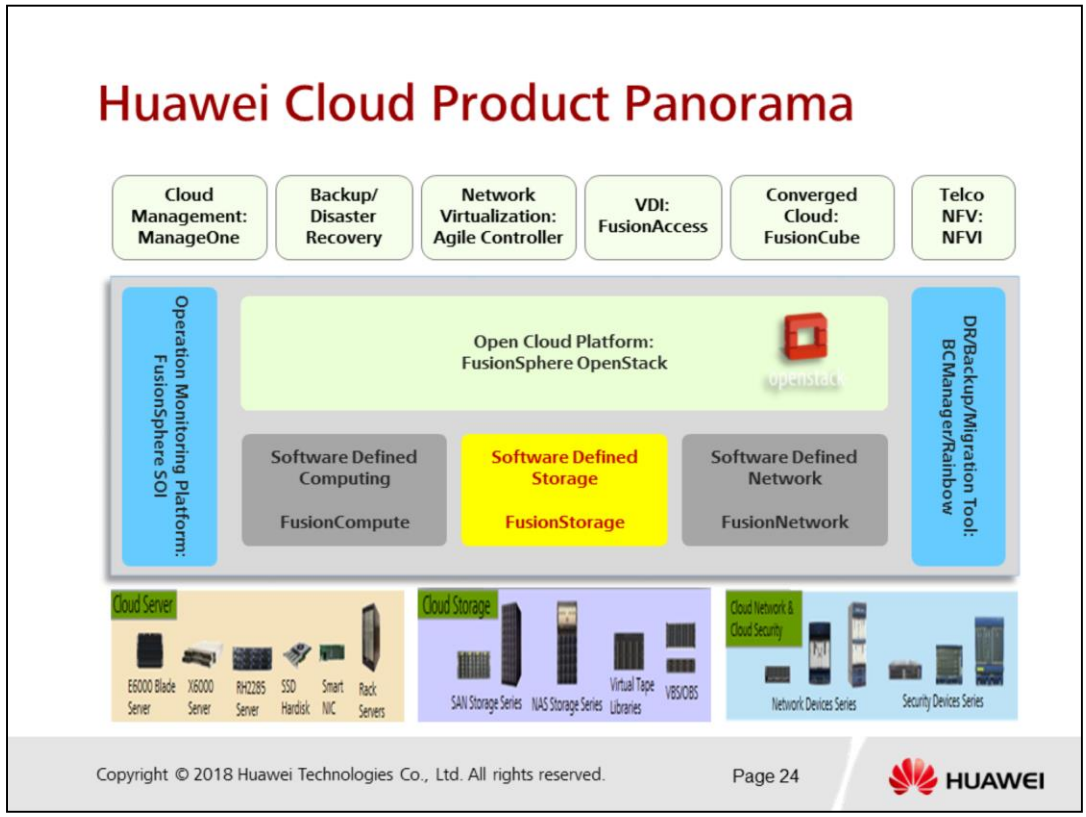
- Converged Storage:
  - Open: Supports deployment of multi vendor virtualized and physical database.
  - High Performance: Supports 4096 nodes, linear performance improvement, supports 10 million IOPS, and supports high performance databases and large capacity cloud resource pools.
  - High Reliability: Multiple backups, and supports up seven 9s reliability level (99.99999%), and has cabinet level reliability.
- Distributed Hash Table (DHT): It is designed for systems that has large number of nodes and the nodes are frequently added or removed.
- A distributed hash table (DHT) is a class of a decentralized distributed system that provides a lookup service similar to a hash table: (key, value) pairs are stored in a DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows a DHT to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures.
- DHTs form an infrastructure that can be used to build more complex services, such as anycast, cooperative Web caching, distributed file systems, domain name services, instant messaging, multicast, and also peer-to-peer file sharing and content distribution systems.

https://t.me/learningnets



- OceanStor DJ is a business-driven storage control software developed by Huawei for cloud data centers. It centrally manages data center storage resources, provides service-driven and automated storage services, improves storage resource utilization and effectiveness of storage service provisioning in cloud environments.
- The core of OceanStor DJ is based on the enhancement of OpenStack related services to realize the unified management of storage resources, on-demand distribution and data protection services. OceanStor DJ decouples applications from underlying storage and breaks the monopoly of legacy devices and application vendors. In cloud based scenarios, storage and data protection capabilities are provided in the form of XaaS, conforming to the storage value chain that is moving towards the direction of software and services.
- From the perspective of control plane and data plane, the value of OceanStor DJ is as following:
    - OceanStor DJ abstracts the storage features from the physical array on the control plane and logically combines multiple physical storage pools with the same or similar capabilities into a virtual storage pool. When a user requests storage resources, the user is allocated resources based on Resource Pool-based Service Level Agreement (SLA) without having to worry about which array on the backend provides storage services for their applications.
    - From the data plane perspective, OceanStor DJ has the ability to integrate all types of data services and the ability to support application access to block storage and file storage. At the same time, it also has the capability to use the unique features of the underlying storage arrays. Hence, OceanStor DJ retains the value added features of the storage arrays such as remote replication without adding to the purchasing cost of the user.

https://t.me/learningnets



- NFV (Network Function Virtualization) refers to the usage of common x86 hardware and virtualization technology to carry out software processing of multiple hardware features in order to lower the cost of purchasing expensive network devices. Through software and hardware decoupling along with the abstraction of features, network devices no longer depends on dedicated hardware. Thus, resources can be fully and flexibly shared, allowing quick development and deployment of new services, achieving automatic deployment based on actual business demands. It also allows elastic expansion and reduction, isolation of faults and self healing of services.

https://t.me/learningnets

Contents

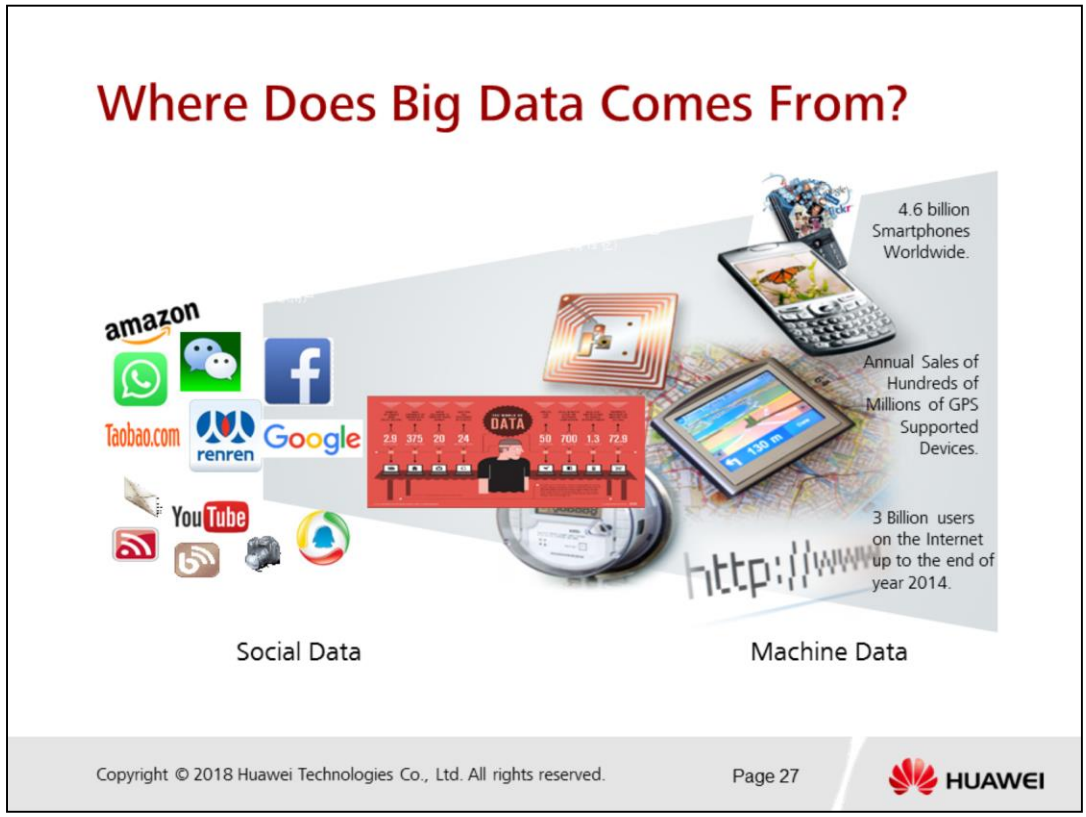
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

https://t.me/learningnets

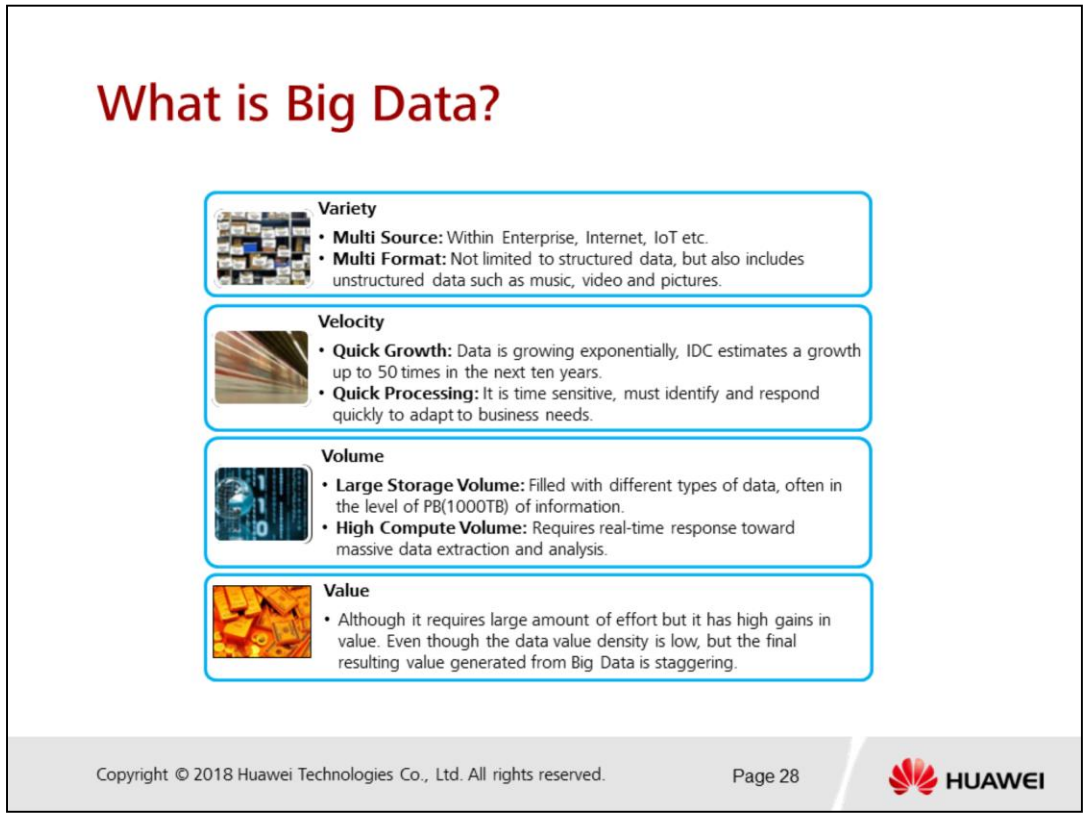
- Big Data Trends: It has overcome the initial stage, and become more and more mature, and it is leading towards the implementation phase.
- 50% of the enterprises has already invested and using Big Data, 33% of the enterprises are planning on how to utilize Big Data. As we see the trends of continual investment on Big Data technologies, Big Data is stepping towards the mature development phase.
- Cloud and Big Data are no longer "new" technologies, but have already become "mainstream" technologies.

https://t.me/learningnets



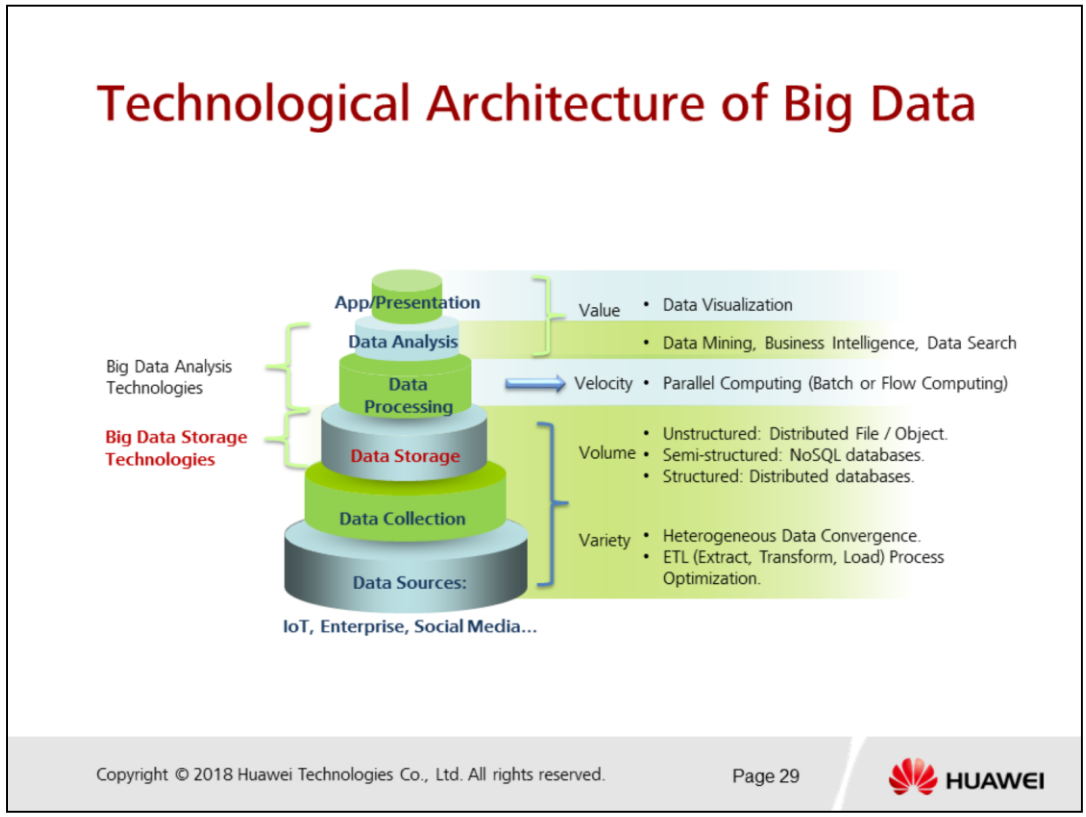
- Digital information created and shared globally increased nine-fold in five years to 3.8ZB (Zettabyte) in 2013.
- CERN (European Organization for Nuclear Research): Large Hadron Collider (LHC) generates 1PB/s worth of data. The LHC is the world's largest and most powerful particle collider, the most complex experimental facility ever built, and the largest single machine in the world.
- SKA(Square Kilometer Array): required storage space of 1EB (Exabyte) in 2015. The Square Kilometre Array (SKA) is a multi-billion dollar international project to build the world's largest radio telescope. Co-located primarily in South Africa and Western Australia, the SKA will be a collection of hundreds of thousands of radio antennas with a combined collecting area equivalent to approximately one million square metres, or one square kilometre.
- Construction of Cloud IDC also generated large concentration of data.
- Facebook: 50TB (Terabyte) of log data is generated daily, and up to 100TB of derived analytical data is generated daily.
- Manufacturing by machines and manufacturing by manual human labor generated massive amounts of data, and on the other hand, construction of centralized data centers also accelerates the centralization of data and data generation.

https://t.me/learningnets



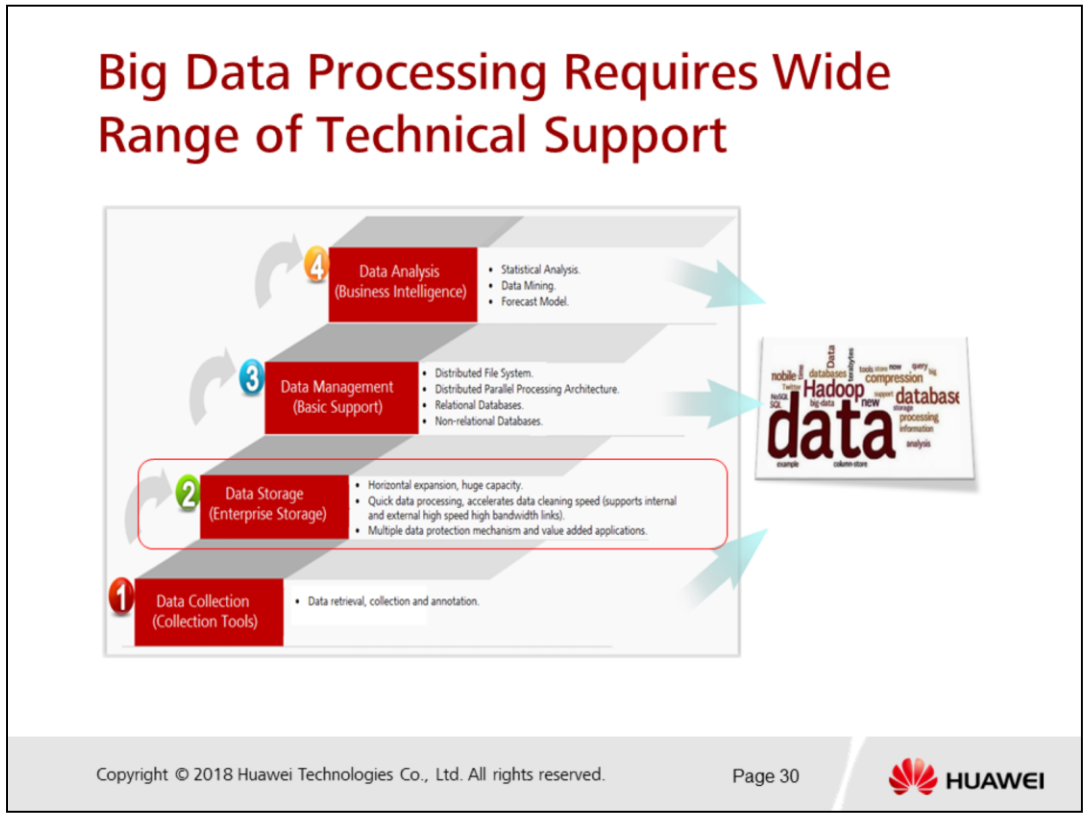
- Wikipedia: Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. The dimensions to big data is known as Volume, Variety, Velocity and Value.
- Gartner: Big Data is a massive, high-growth and diversified information asset that requires new processing models for greater decision-making, insight, and process optimization capabilities.
- The essence of Big Data: It is the reflection of the physical world in the digital world. For example, the flow of people in annual holidays will be recorded in the digital world.
- The logic of real world phenomena can also be found through big data analysis. For example, when a storm comes, we can see that the seagulls fly low. Through analysis, it is found that the seagulls fly low due to the appearance of many fishes on the surface of the sea which they could easily prey upon. So, why does the fishes swims near to the surface of the sea? Apparently, when the storm is coming, the water pressure in the sea increases, and fishes swim near to the surface of the sea where it is more convenient for them to breathe. These are all the relationship behind the phenomenon that can be figured out through big data analysis.
- Big data doesn't just refers to new data that are collected, in our society, the most valuable data still lies within the data that was accumulated by the enterprise over the years. Thus, data that was generated in the core systems of traditional data management is still relevant and valuable. It may not be the biggest amount of data or the latest or trendiest data, but it is the data that holds the highest commercial value to the enterprise itself.

https://t.me/learningnets



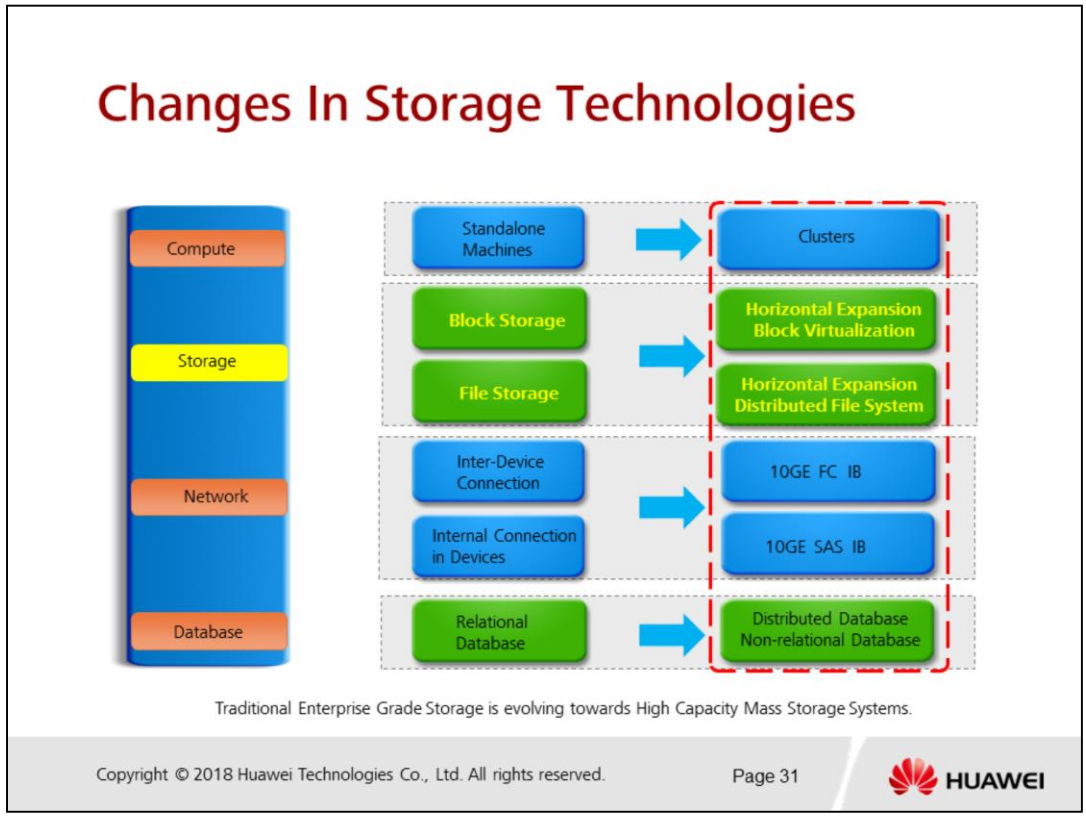
- From the perspective of the challenges faced by big data in terms of capacity, data diversity, processing speed and value mining, big data technologies cover a wide range of technologies ranging from mass storage and processing of data to applications that visualize those data, including convergence of heterogeneous data sources, massive distributed file system, NoSQL database, parallel computing framework, live stream data processing and data mining, business intelligence and data visualization.
- A typical big data processing system can be divided into 5 layers: Data Collection, Data Storage, Data Processing, Data Analysis and Application or Presentation of Data. The architecture of these layers can be seen as the diagram shown above.
- The data types and models of Big Data:
  - Unstructured Data: refers to data that not easily presented in the form of database or 2D logical charts, including all the formats of office files, text, pictures, XML, HTML, charts, images, music, and video.
  - Semi-structured Data: refers to the data that sits between structured data (such as relational database or object oriented database data) and unstructured data (such as sound and images). HTML files can be considered as a semi-structured data, as it is self descriptive, as the content and structure of the data are mixed together without clear distinction.
  - Structured Data: refers to data that can be stored within databases and can be represented or visualized in the form of 2D logical chart.

https://t.me/learningnets



- The process of processing big data is basically the same as the process of general data processing, which includes the four process of data collection, data storage, data management and data analysis.
- The data collection process refers to collecting the required data for subsequent analysis through various sources, methods and tools. In Big Data scenarios, what kind of data needs to be collected? By what means should those data be collected? What methods should be used to collect those data? Which tool should be used for more efficient data collection? These are some of the questions that needs to be put under considerations during this process.
- The data storage process refers to the transmission and storage of data that was collected in the earlier process. With the exponential growth of data, traditional storage systems are no longer able to meet the requirements of Big Data storage. Thus, new technologies are needed to achieve the storage requirements of Big Data.
- The data management process can be considered as an extension to the data storage process. The data is further processed from the basis of the stored data, to realize much finer data segmentation and provides directly available metadata that improves the data analysis efficiency for subsequent data analysis process.
- The data analysis process refers mainly to the utilization of analysis methods, models, and tools to analyze the data and getting a final analysis result that satisfies the common analytical needs. Deeper mining based on the data that was queried and analyzed earlier, allows us to meet higher level of data analysis requirements.

https://t.me/learningnets

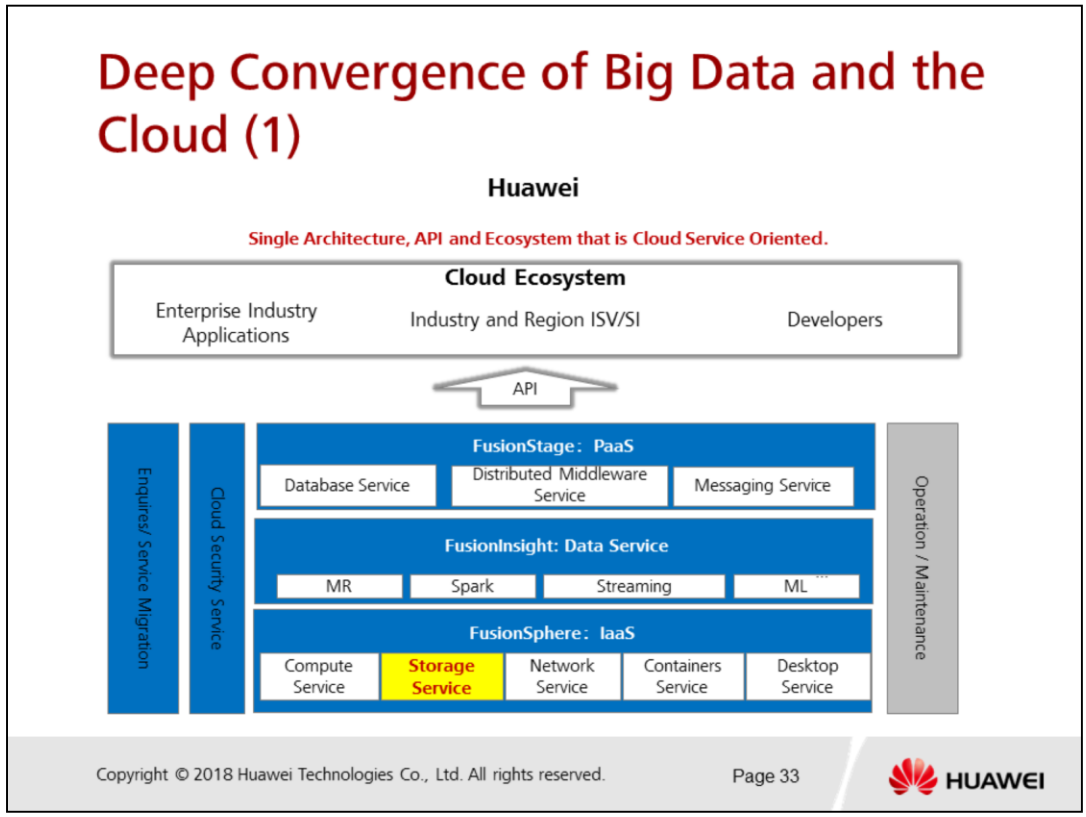


- Era of Big Data:
  - Compute: Developing towards the direction of clustering.
  - Storage: Both block and file storage are developing towards horizontal expansion, virtualization and rich software interfaces for external connections. However, file storage devices has higher horizontal expansion capabilities, and its hardware devices usually expands up to hundreds of nodes. File system also develops from local file system towards clustered file system and distributed file systems.
  - Network: No matter if it is the connection within the Internet or within internal networks, it is developing towards a direction of higher speeds, lower protocol overhead and more efficiency.
  - Database: The development trends moves from relational database towards distributed database and non-relational database (such as In Memory Database).
- The changes in these technologies brought upon new points of opportunities:
  - Database Revolution: Relational Database → Non-relational Database → Hybrid Database.
  - File System Revolution: Local File System → Clustered File System → Distributed File System.

https://t.me/learningnets

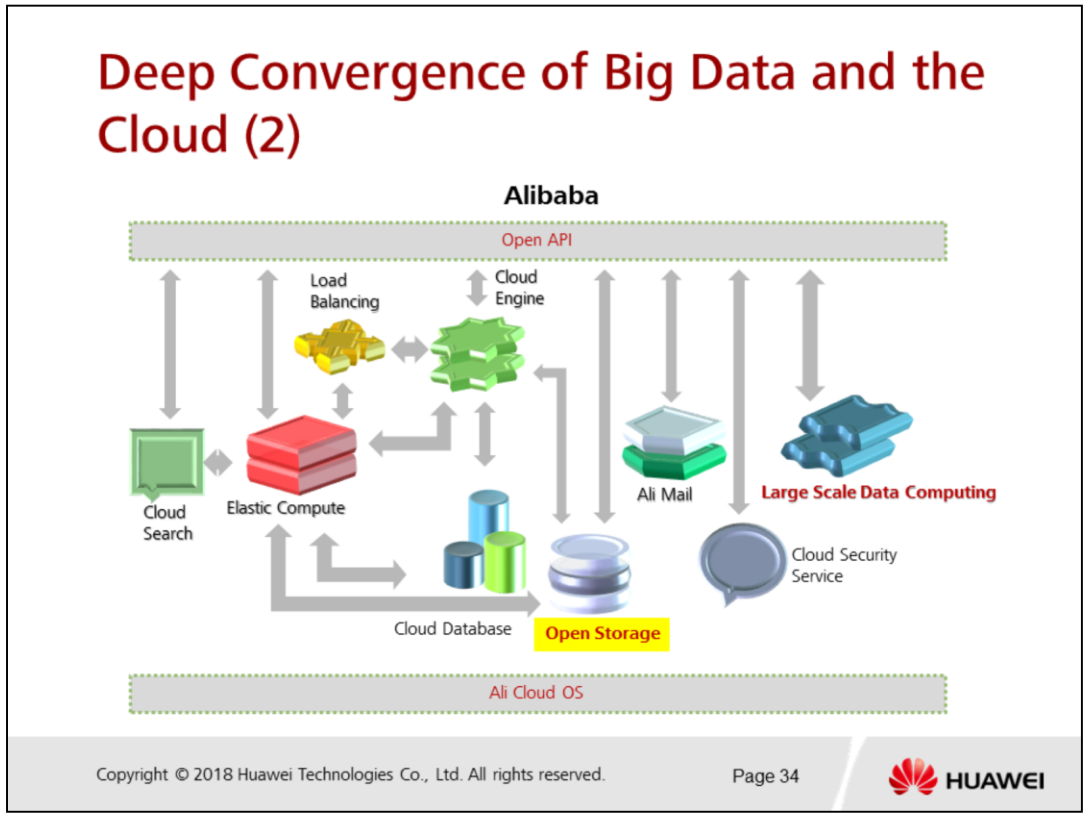
- Challenges faced by traditional enterprise grade storage:
    - Traditional SAN that has dual controller or multi controller architecture, is unable to store and handle data in the scale of Petabytes (PB).
    - Traditional NAS that has complex volume management and unbalanced capacity allocation, causes resource wastage when handling data in the scale of PB level.
- Features of Mass Storage Systems:
    - Single Unified File System which makes management much simpler.
    - Simple configuration, quota management, capacity allocation and has high disk utilization rate.
    - Large scale horizontal expansion and scalability.

https://t.me/learningnets



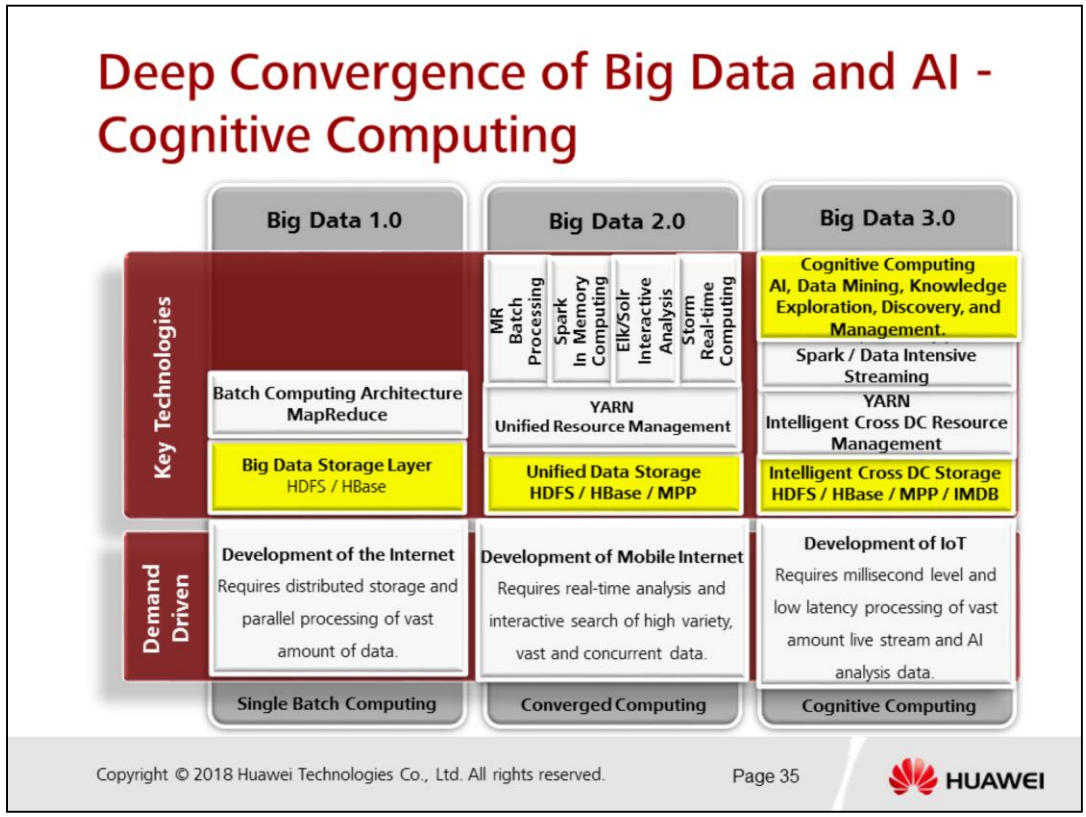
- Huawei focuses on its identity as the technology provider in the Big Data industry, where they provide the physical infrastructure, hardware devices, cloud platform and big data analysis platforms for enterprises in the market looking to build their own Big Data systems.
- The deep convergence of the cloud and big data can be further explained as follows:
- IaaS which is a part of the cloud provides the data storage and analysis services for Big Data, and this layer mainly corresponds and faces the personnel with the role of data administrators. Data administrators manages the vast amounts of data in terms of actual storage and data processing for Big Data.
- PaaS which is also a part of the cloud provides the development services for Big Data, where this layer mainly corresponds and faces the personnel with the role of developers. The developers create applications or software systems to utilize, process and visualize the data to form meaningful insights and value.
- Industry solutions, Big Data storage, Big Data analysis, and Big Data Applications are often represented in the form of SaaS, hence, Big Data can also be considered as one of the basic services offered in the cloud.

https://t.me/learningnets



- Big Data requires a certain technical service level to be fully utilized in a correct manner, but this issue is solved and covered by providing products and solution trainings, Huawei Big Data certifications, industry talent training programs and industry recognized certifications.
- Big data processing is inseparable from Cloud technology. Cloud provides a flexible and scalable infrastructure supporting environment for big data and is an efficient mode of data services. On the other hand, Big data provides new business value for Cloud. In overall, emerging computing trends such as Cloud, the Internet of Things, and the Mobile Internet are both areas that generate vast amounts of data which requires the analytical processing of Big Data.
- ISV refers to Independent Software Vendors, while SI refers to System Integrators where both of them are important partners that assist in building a cloud data center for enterprises and integrate all their existing business systems and even provide new systems or application that helps in the business operation of the enterprises.

https://t.me/learningnets



- AI refers to Artificial Intelligence.
- MapReduce (MR) is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.
- Spark is the leading platform for large-scale SQL, batch processing, stream processing, and machine learning.
- "ELK" is the acronym for three open source projects: Elasticsearch, Logstash, and Kibana. Elasticsearch is a search and analytics engine. Logstash is a server-side data processing pipeline that ingests data from multiple sources simultaneously, transforms it, and then sends it to a "stash" like Elasticsearch. Kibana lets users visualize data with charts and graphs in Elasticsearch.
- Solr is an open source enterprise search platform. Its major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is designed for scalability and fault tolerance. Solr is widely used for enterprise search and analytics use cases and has an active development community and regular releases.

https://t.me/learningnets

- HDFS (Hadoop Distributed File System) is a distributed file system that handles large data sets running on commodity hardware that can be used for scaling a single cluster to hundreds (and even thousands) of nodes.
- HBase is a column-oriented database management system that runs on top of Hadoop Distributed File System (HDFS). It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL. In fact, HBase isn't a relational data store at all.
- MPP (massively parallel processing) is the coordinated processing of a program by multiple processors working on different parts of the program. Each processor has its own operating system and memory. MPP speeds the performance of huge databases that deal with massive amounts of data.
- YARN (Yet Another Resource Negotiator) is a large-scale, distributed clustering platform that helps to manage resources and schedule tasks in Hadoop Big Data scenarios.
- Apache Hadoop is an open source software framework that can be installed on a cluster of commodity machines so the machines can communicate and work together to store and process large amounts of data in a highly distributed manner.
- Apache Storm is a free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data. Storm is extremely fast, with the ability to process over a million records per second per node on a cluster of modest size. Enterprises harness this speed and combine it with other data access applications in Hadoop to prevent undesirable events or to optimize positive outcomes in their Big Data scenarios.

https://t.me/learningnets

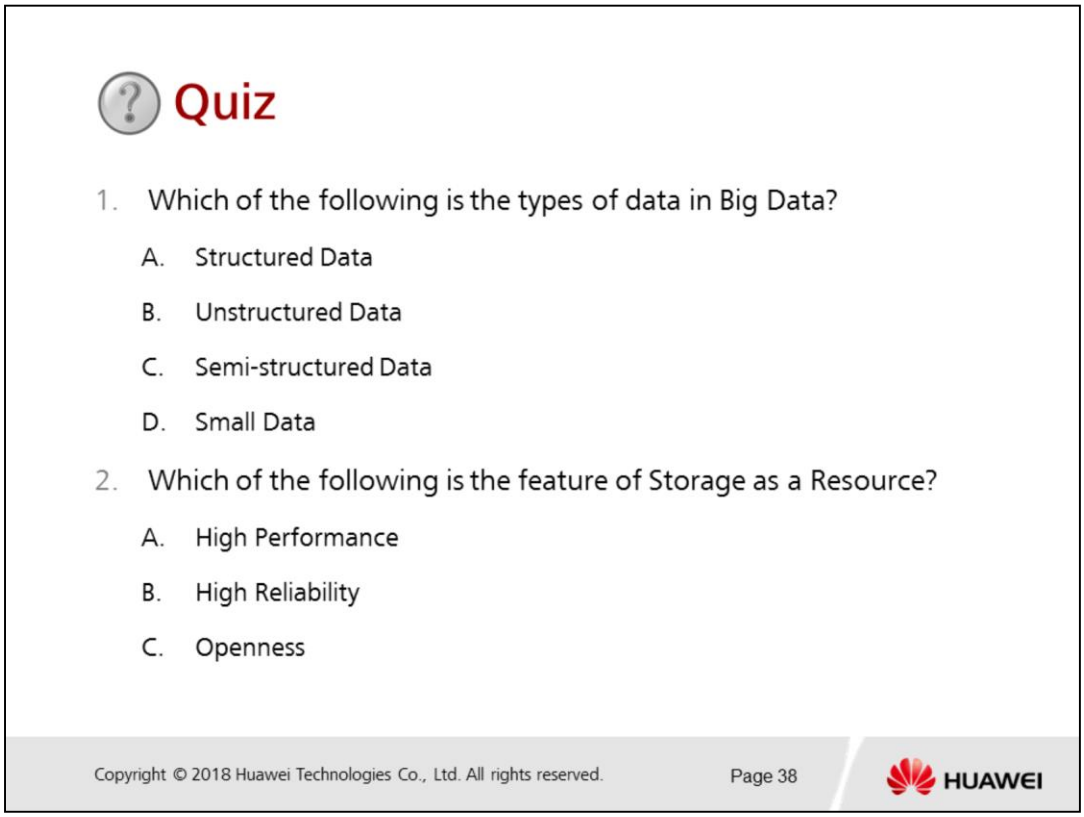
Summary

- This module mainly introduced:
  - The Development Trends of ICT Architecture.
  - Concepts of the Cloud, Big Data and AI.
  - Storage Technologies and Its Application in the Cloud.
  - Storage Technologies and Its Application in AI and Big Data.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

https://t.me/learningnets



- Answers:
  - ABC.
  - ABC.

https://t.me/learningnets

Thank You

www.huawei.com

https://t.me/learningnets