



Big Data and Internet Thinking

Chentao Wu

Associate Professor

Dept. of Computer Science and Engineering

wuct@cs.sjtu.edu.cn



上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

Download lectures

- <ftp://public.sjtu.edu.cn>
- User: wuct
- Password: wuct123456
- <http://www.cs.sjtu.edu.cn/~wuct/bdit/>

Schedule

- lec1: Introduction on big data, cloud computing & IoT
- lec2: Parallel processing framework (e.g., MapReduce)
- lec3: Advanced parallel processing techniques (e.g., YARN, Spark)
- lec4: Cloud & Fog/Edge Computing
- lec5: Data reliability & data consistency
- lec6: Distributed file system & objected-based storage
- lec7: Metadata management & NoSQL Database
- lec8: Big Data Analytics

Final Grade

- Attendance 20%
- Reports & Projects 80%
 - Reports and Projects will be checked by TA.

Collaborators



1

Introduction to Big Data



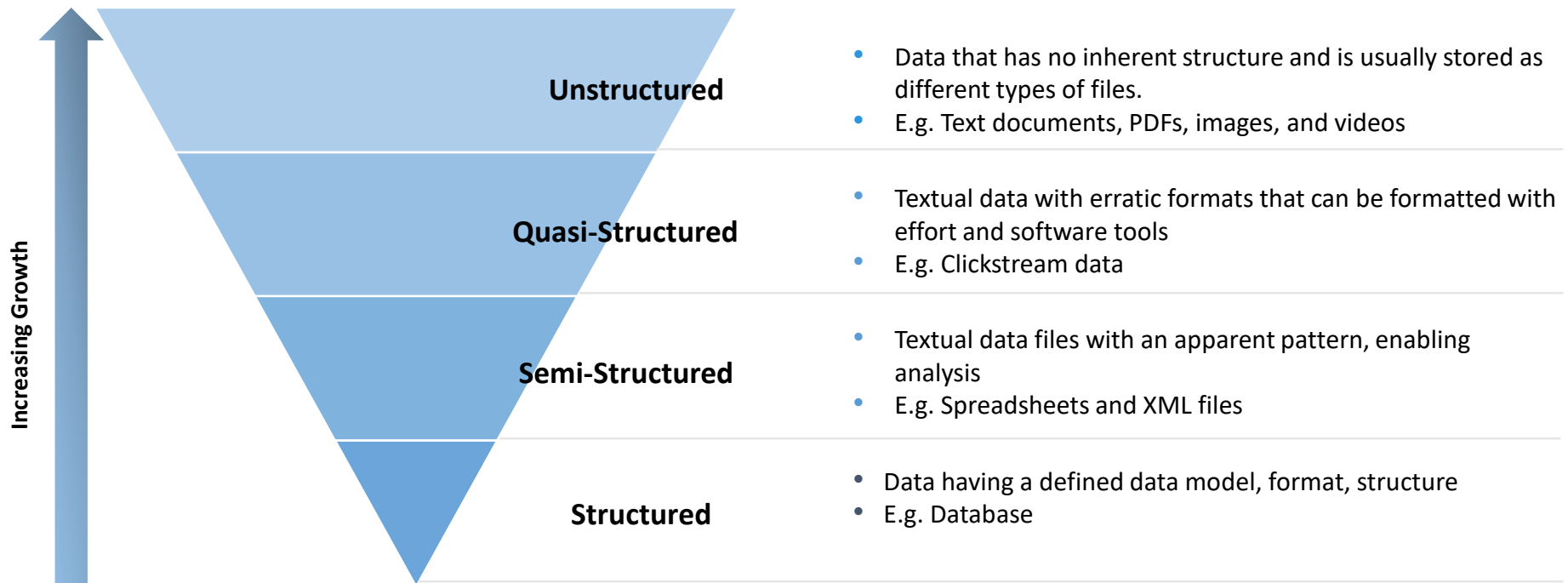
Big Data Definition

- No single standard definition...

“**Big Data**” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...

Types of Data

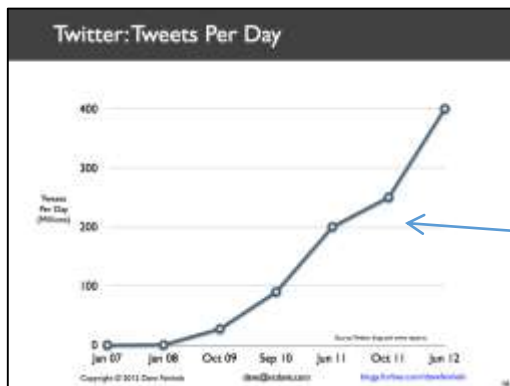
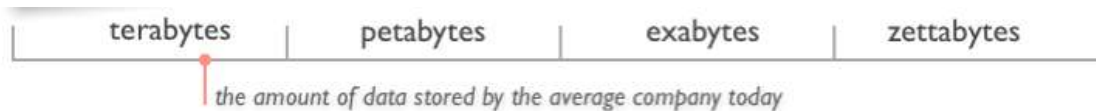
- Structured
- Semi-Structured/Quasi-Structured/Unstructured



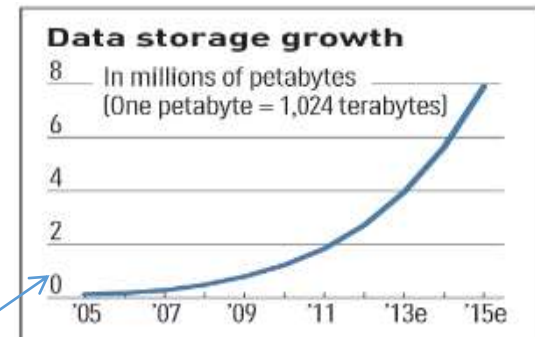
Characteristics of big data (1-Scale: Volume)

- **Data Volume**

- 44x increase from 2009 2020
- From 0.8 ZettaBytes to 44ZB
- Data volume is increasing exponentially



The Digital Universe 2009-2020



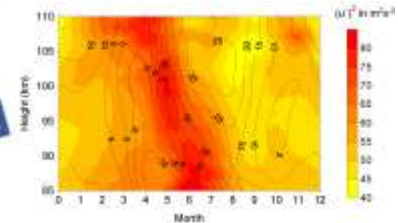
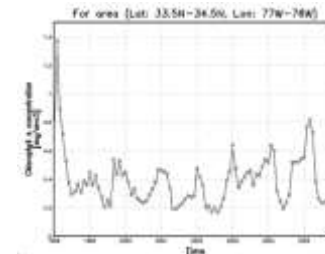
Exponential increase in
collected/generated data

Characteristics of big data

(2-Complexity: Varsity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc...
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

To extract knowledge → all these types of data need to be linked together



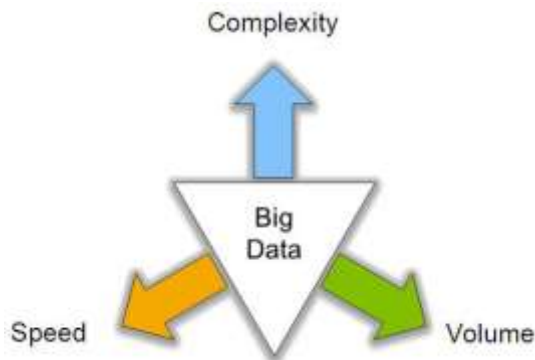
Characteristics of big data

(3-Speed: Velocity)

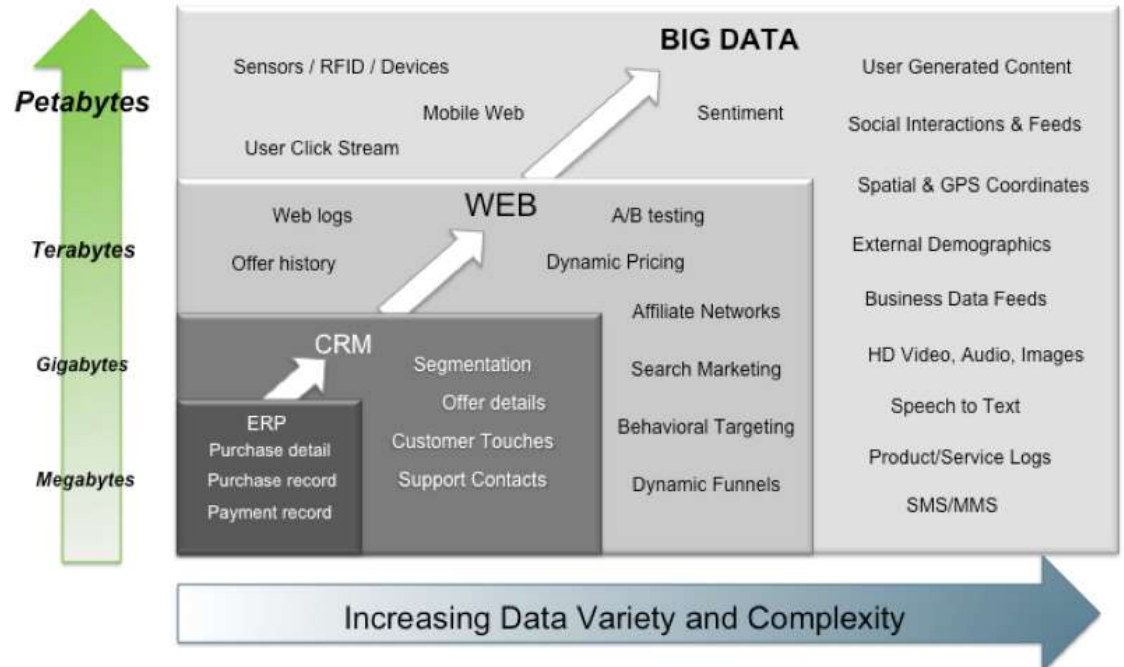
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- **Examples**
 - **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions right now for store next to you
 - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction



Big Data (3Vs)

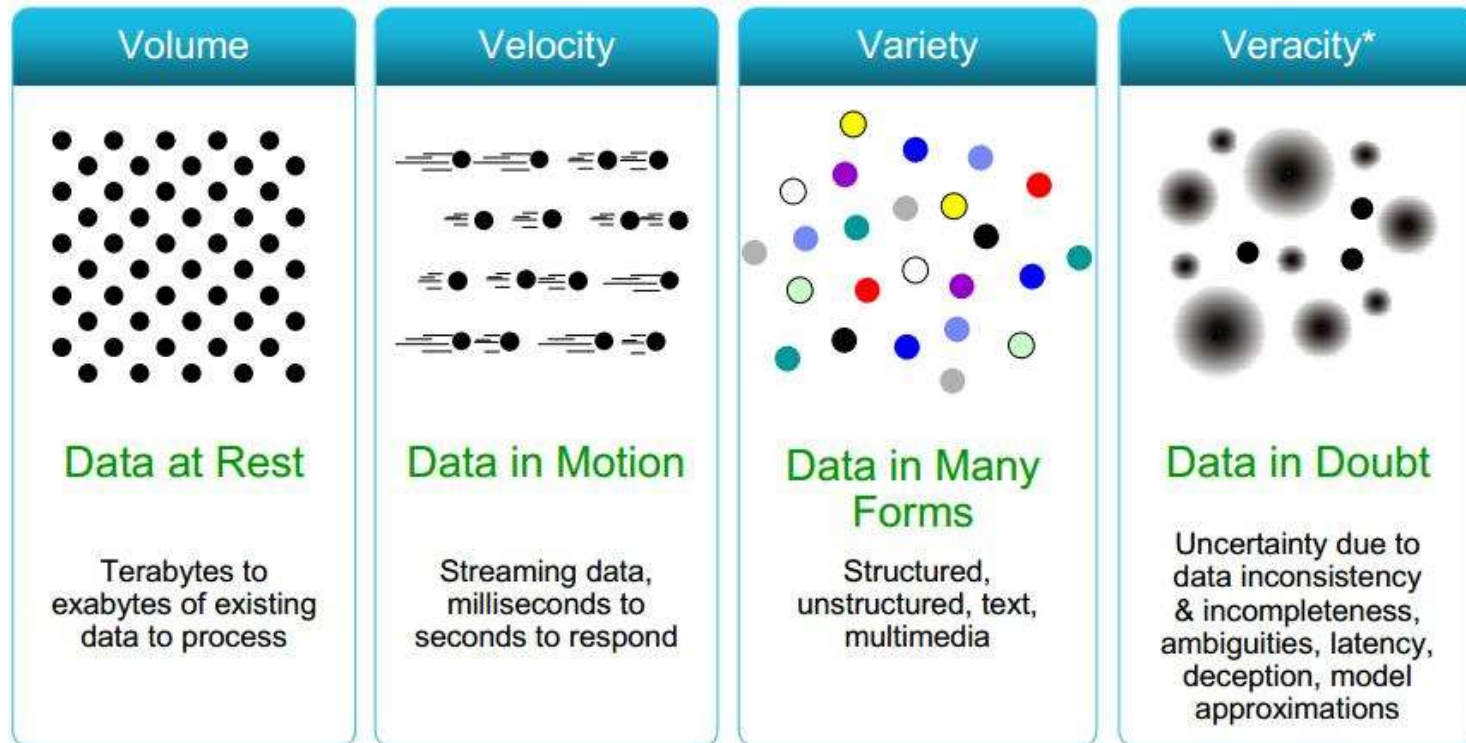


Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

Big Data (4Vs)



Big Data (5Vs/6Vs)



Volume

- Massive volumes of data
- Challenges in storage and analysis



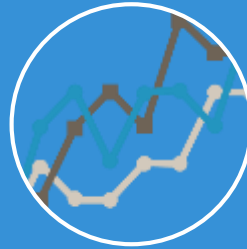
Velocity

- Rapidly changing data
- Challenges in real-time analysis



Variety

- Diverse data from numerous sources
- Challenges in integration, and analysis



Variability

- Constantly changing meaning of data
- Challenges in gathering and interpretation



Veracity

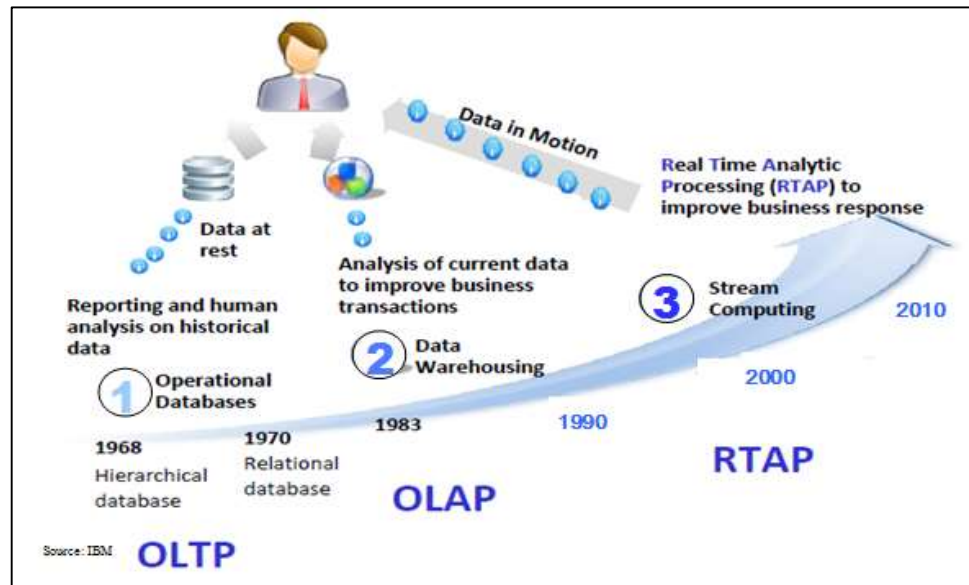
- Varying quality and reliability of data
- Challenges in transforming and trusting data



Value

- Cost-effectiveness and business value

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Who's Generating Big Data



Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**

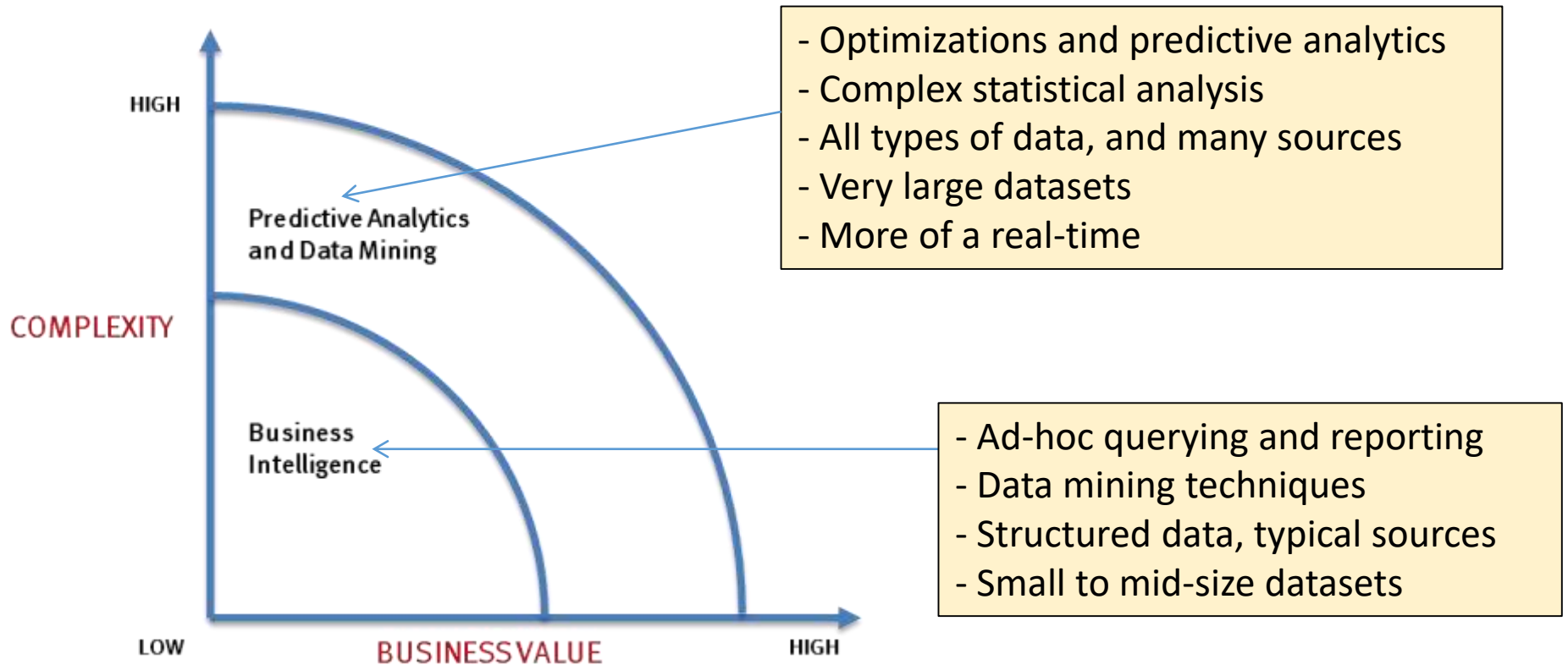
Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data

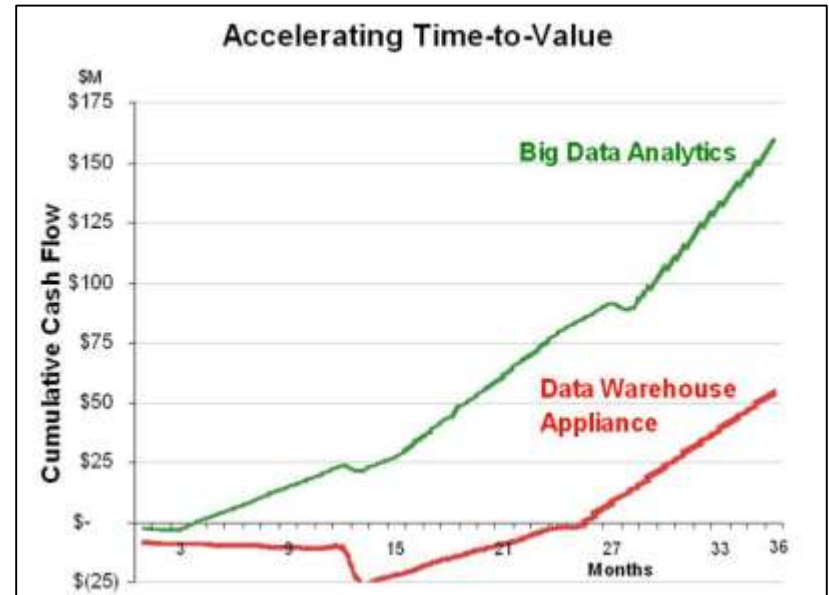


What's driving Big Data

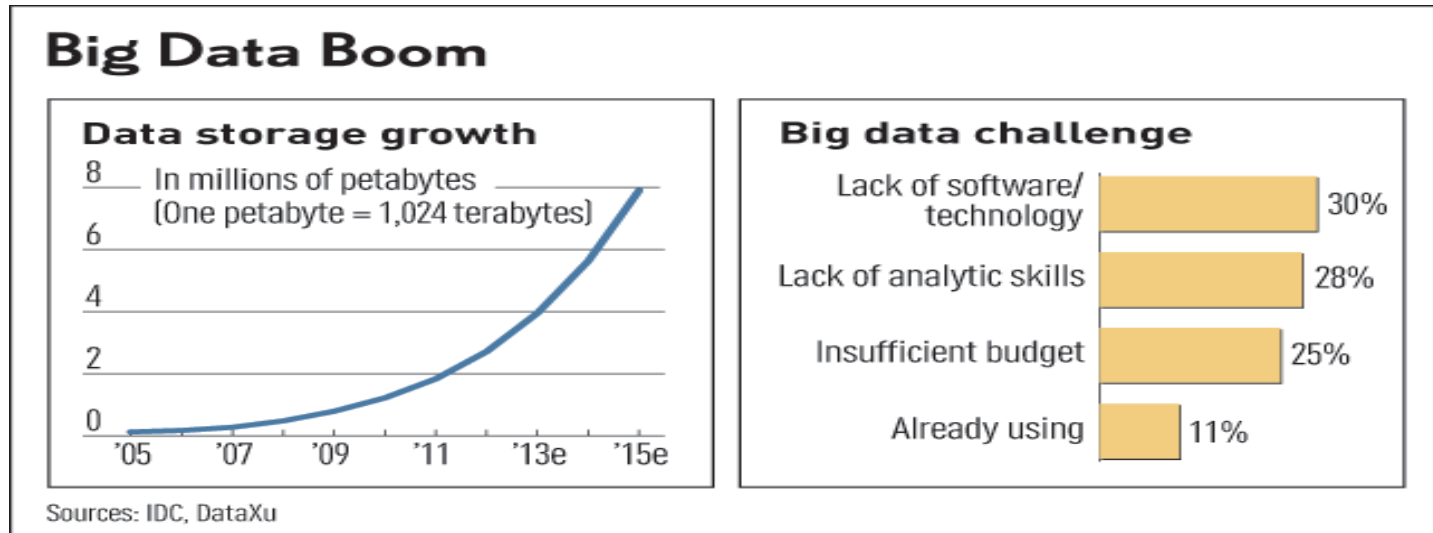


Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



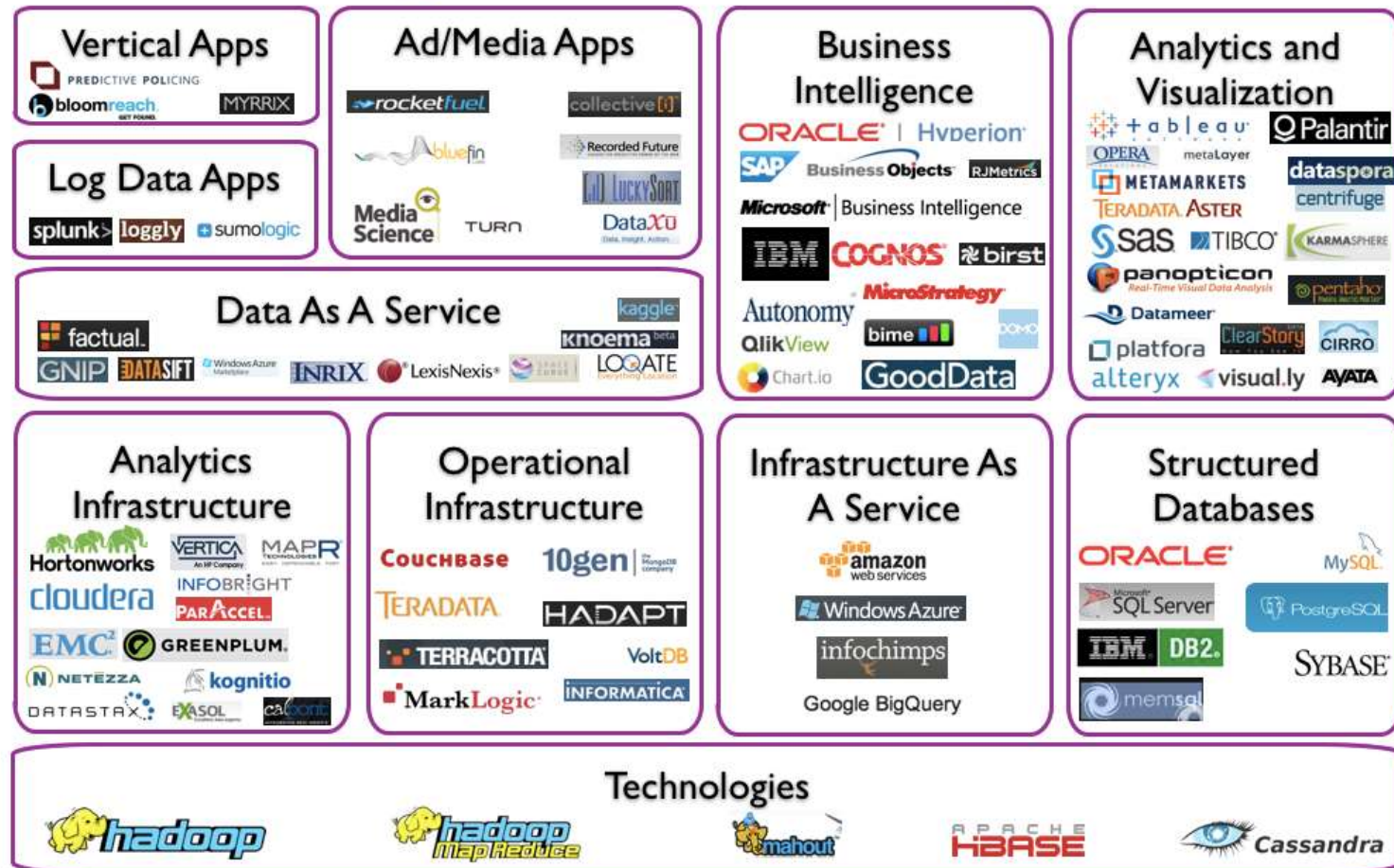
Challenges in Handling Big Data



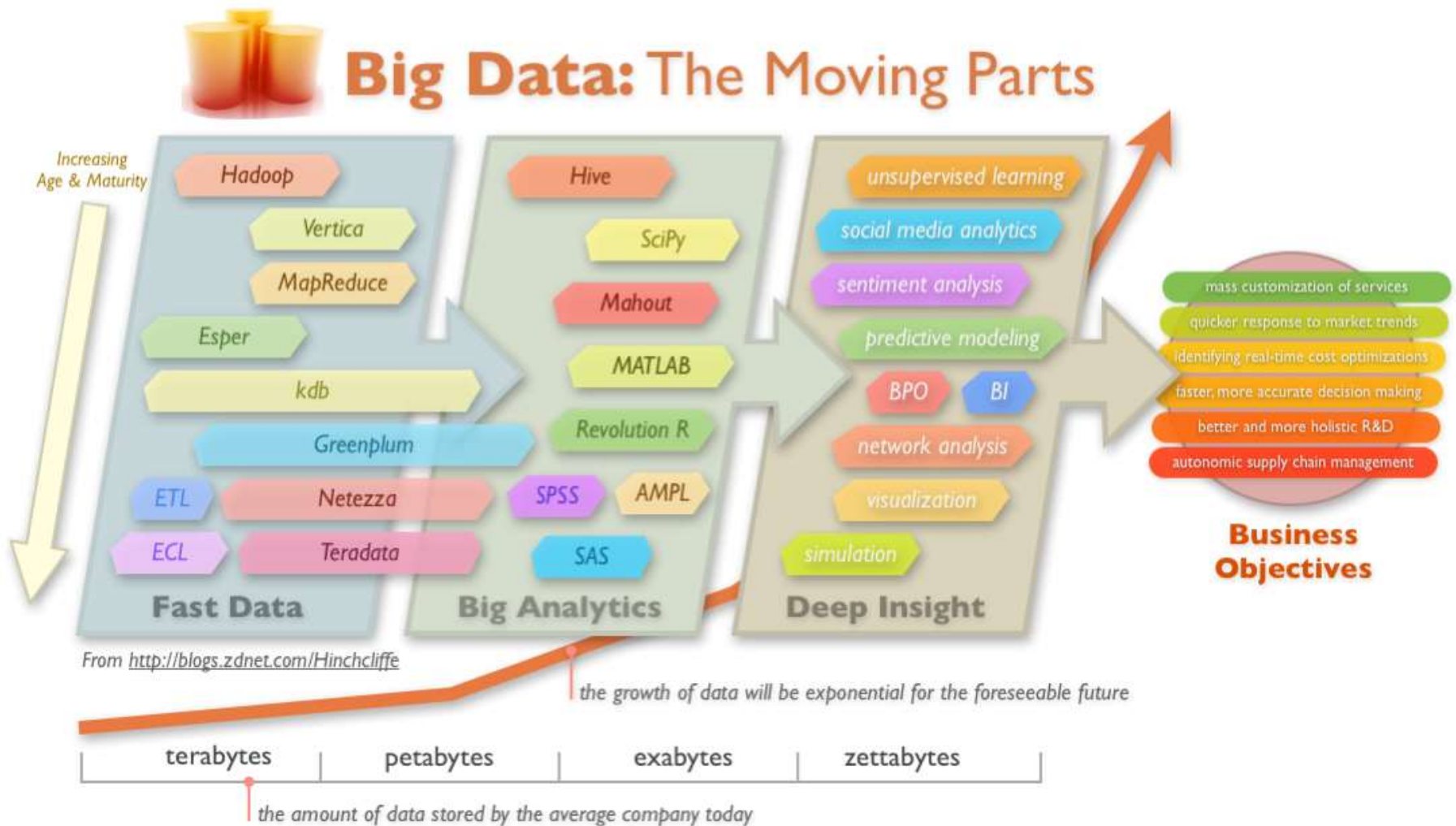
- **The Bottleneck is in technology**
 - New architecture, algorithms, techniques are needed
- **Also in technical skills**
 - Experts in using the new technology and dealing with big data

Big Data Landscape

Big Data Landscape



Big Data Technology



2

Introduction to Cloud Computing



What is Cloud Computing?

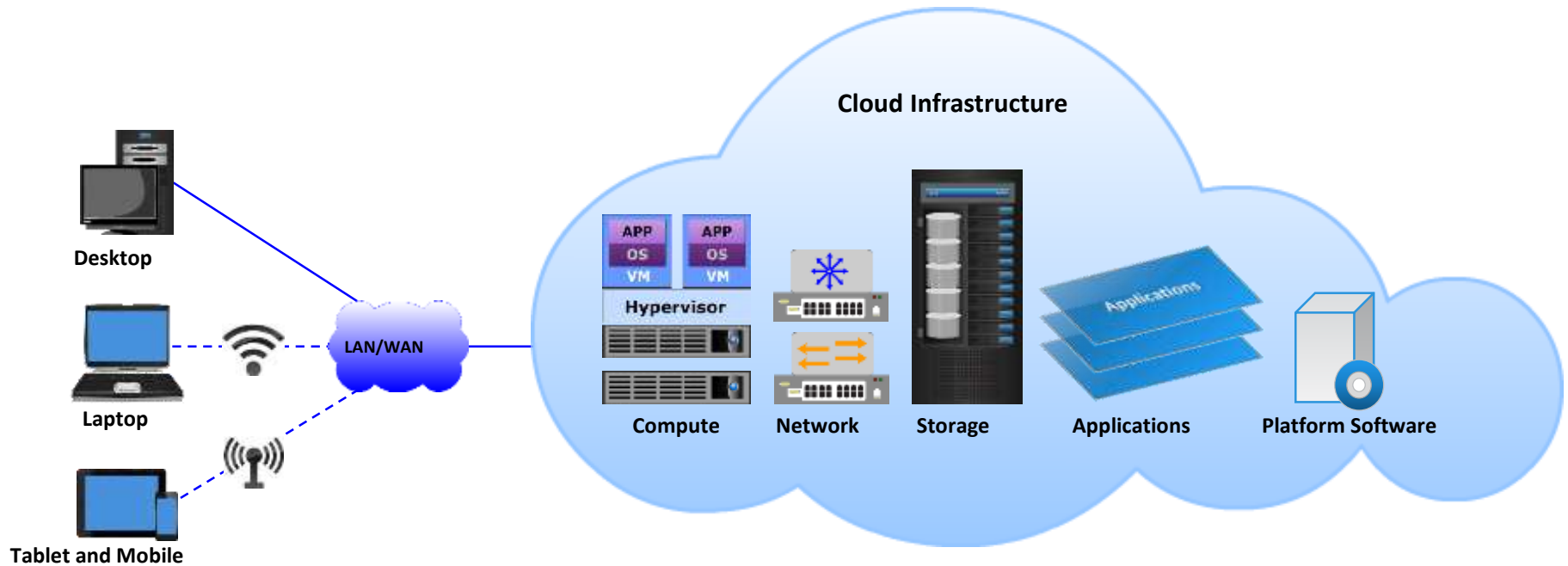
Cloud Computing

A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources, (e.g., servers, storage, networks, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

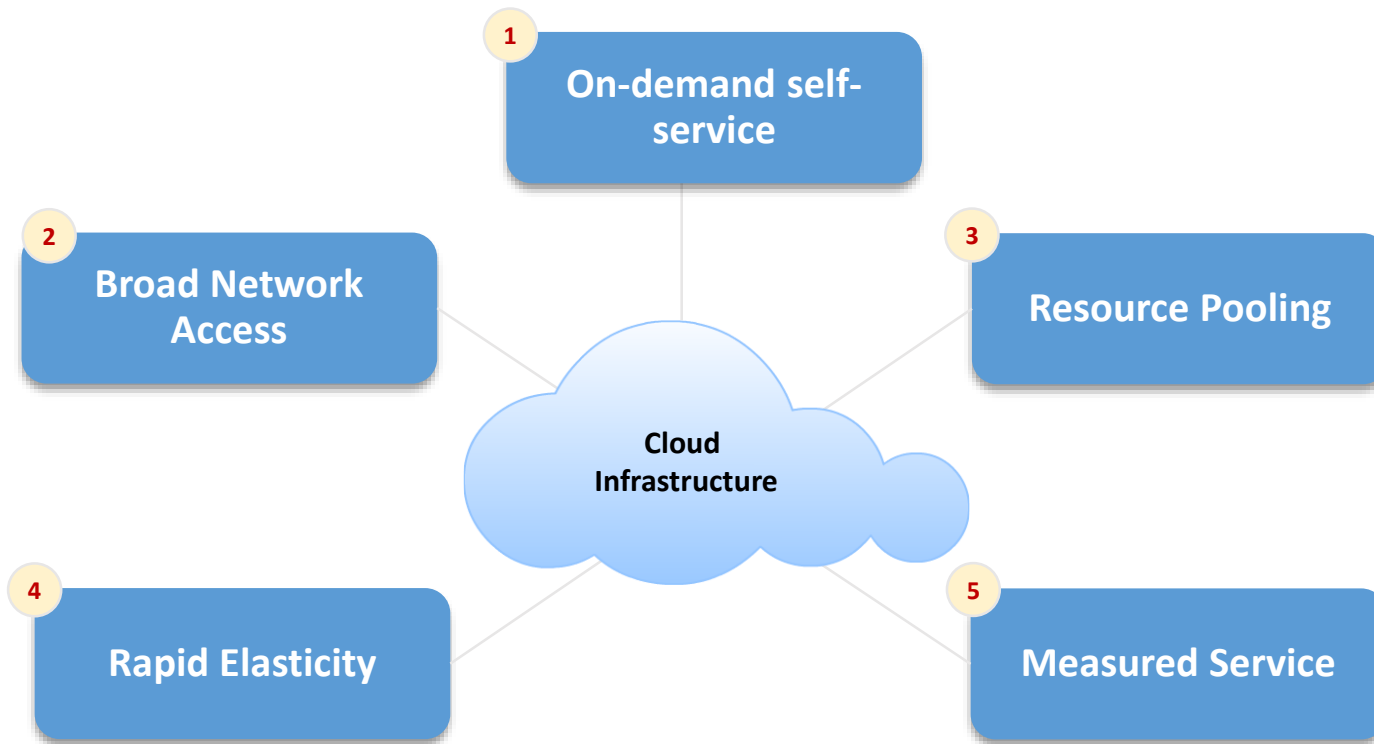
– U.S. National Institute of Standards and Technology, Special Publication 800-145

- A cloud is a collection of network-accessible hardware and software resources
 - Consists of IT resource pools deployed in data centers
- Cloud model enables consumers to hire IT resources as services

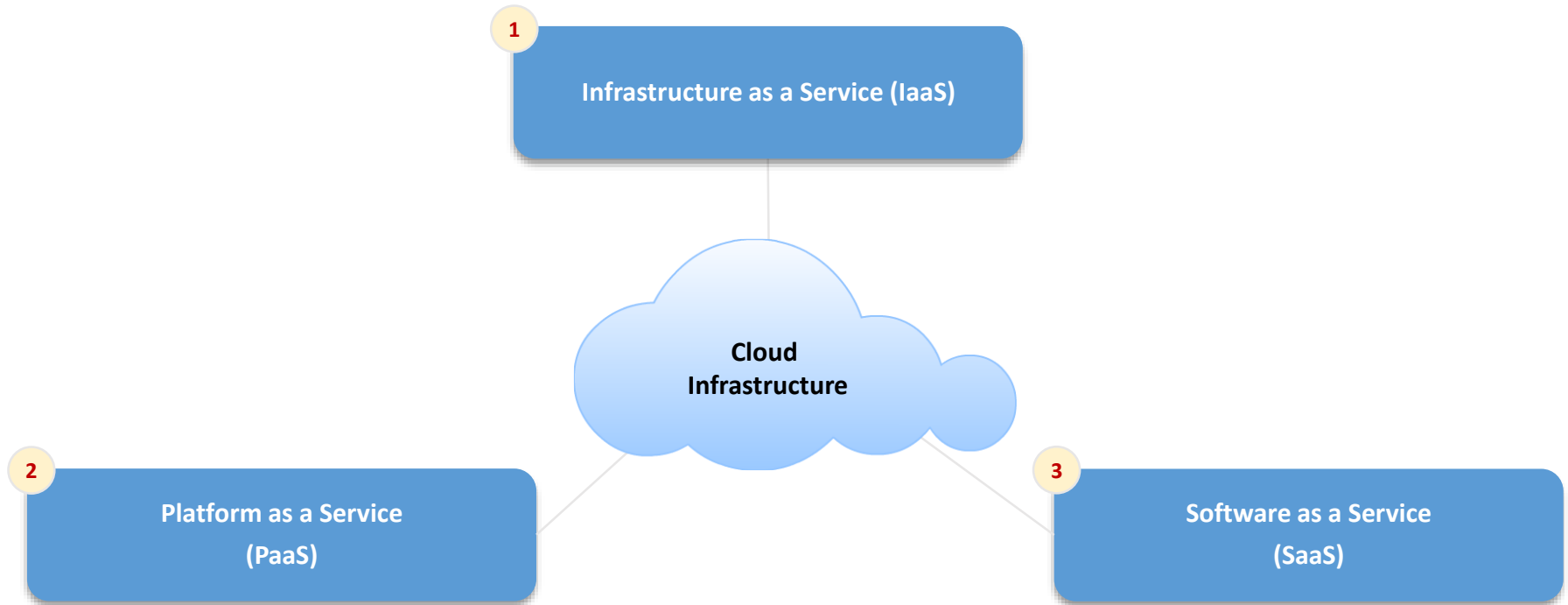
What is Cloud Computing? (Cont'd)



Essential Cloud Characteristics



Cloud Service Models

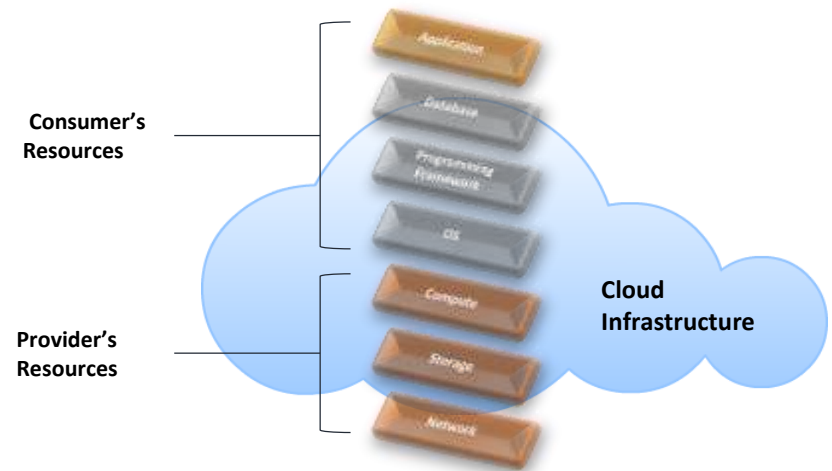


Infrastructure as a Service

Infrastructure as a Service

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components, (e.g., host firewalls).

– U.S. National Institute of Standards and Technology, Special Publication 800-145

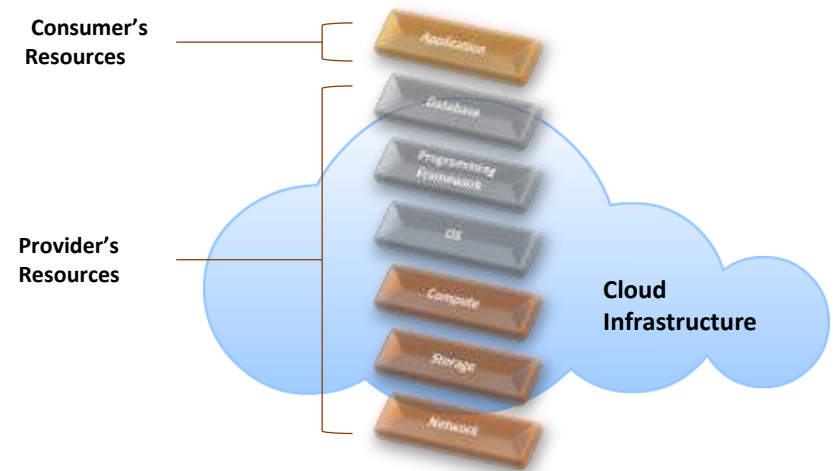


Platform as a Service

Platform as a Service

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

– U.S. National Institute of Standards and Technology, Special Publication 800-145

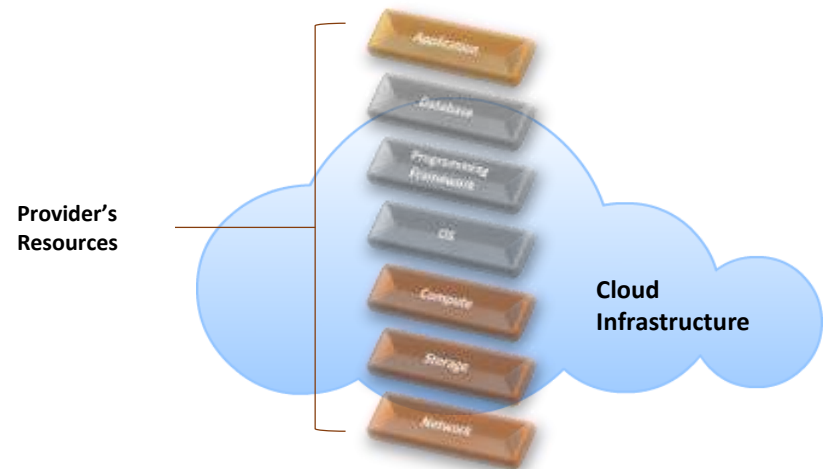


Software as a Service

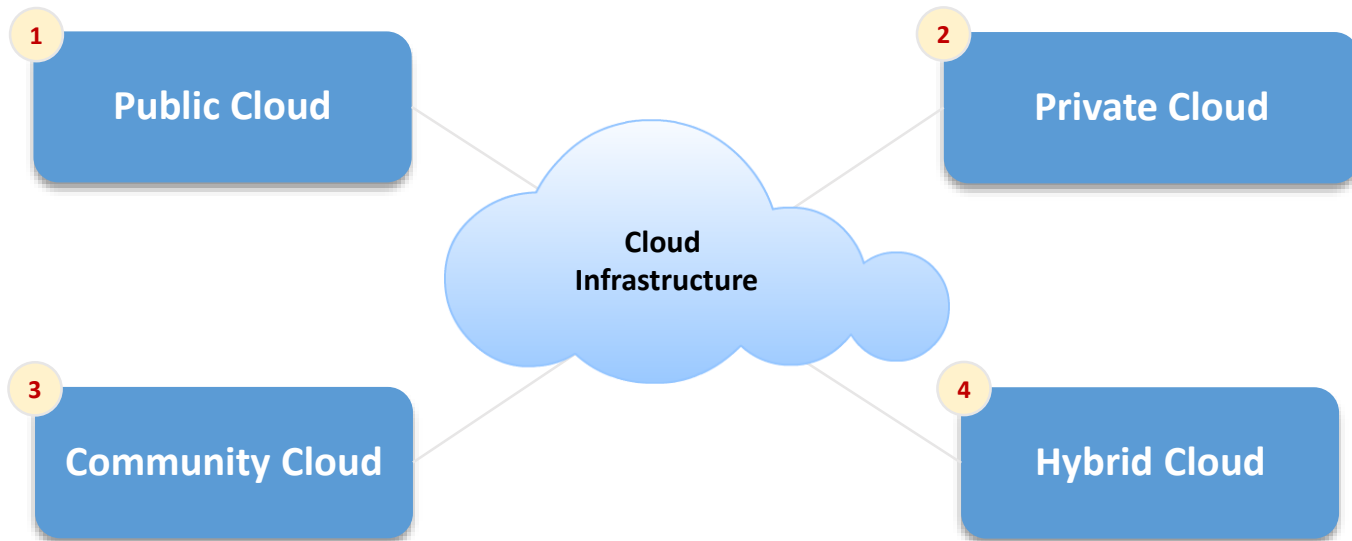
Software as a Service

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser, (e.g., web-based email, or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

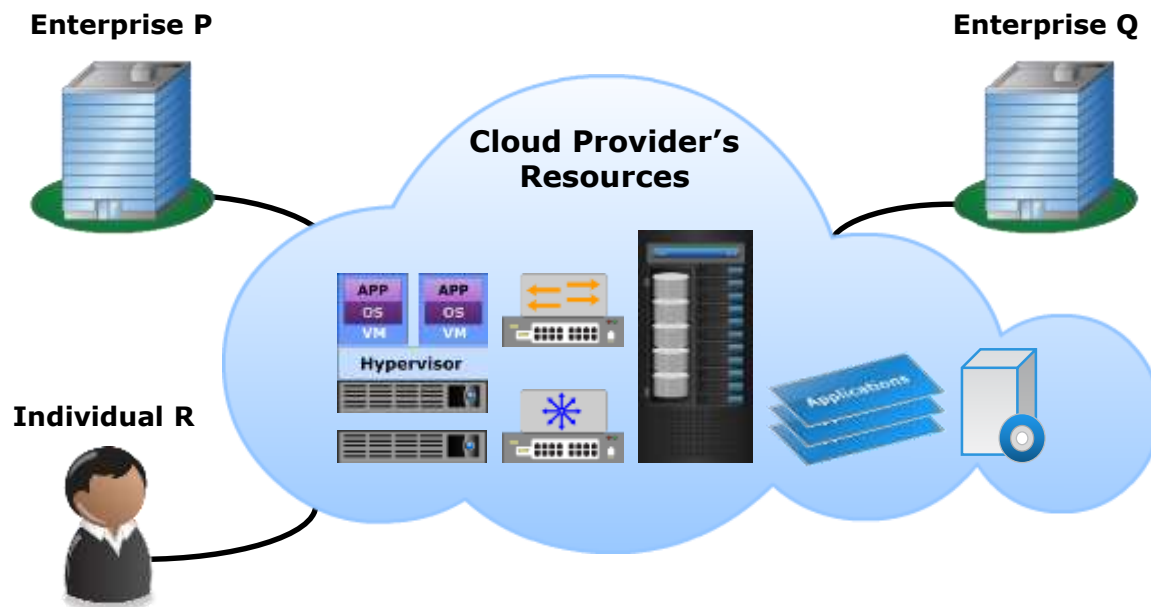
– U.S. National Institute of Standards and Technology, Special Publication 800-145



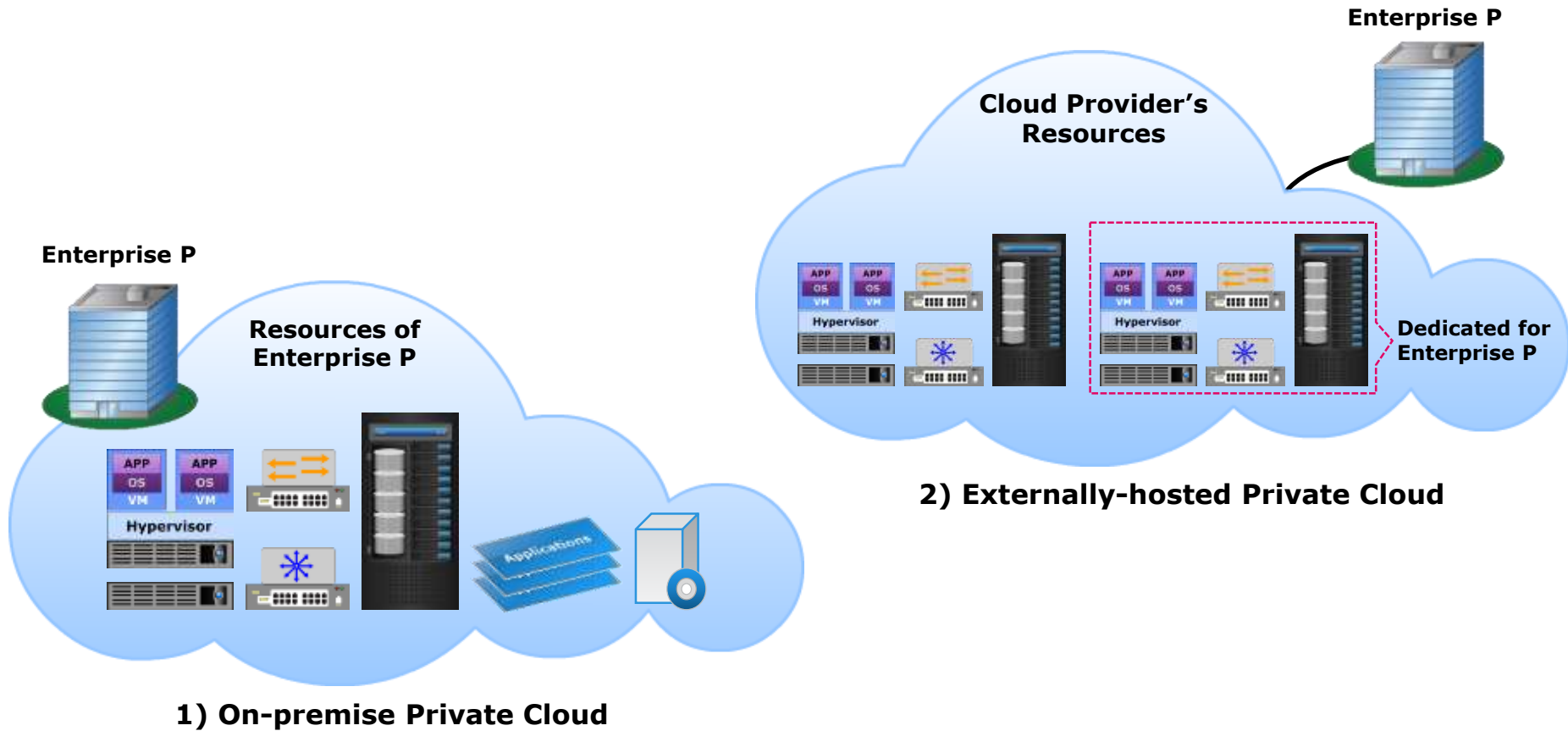
Cloud Deployment Models



Public Cloud

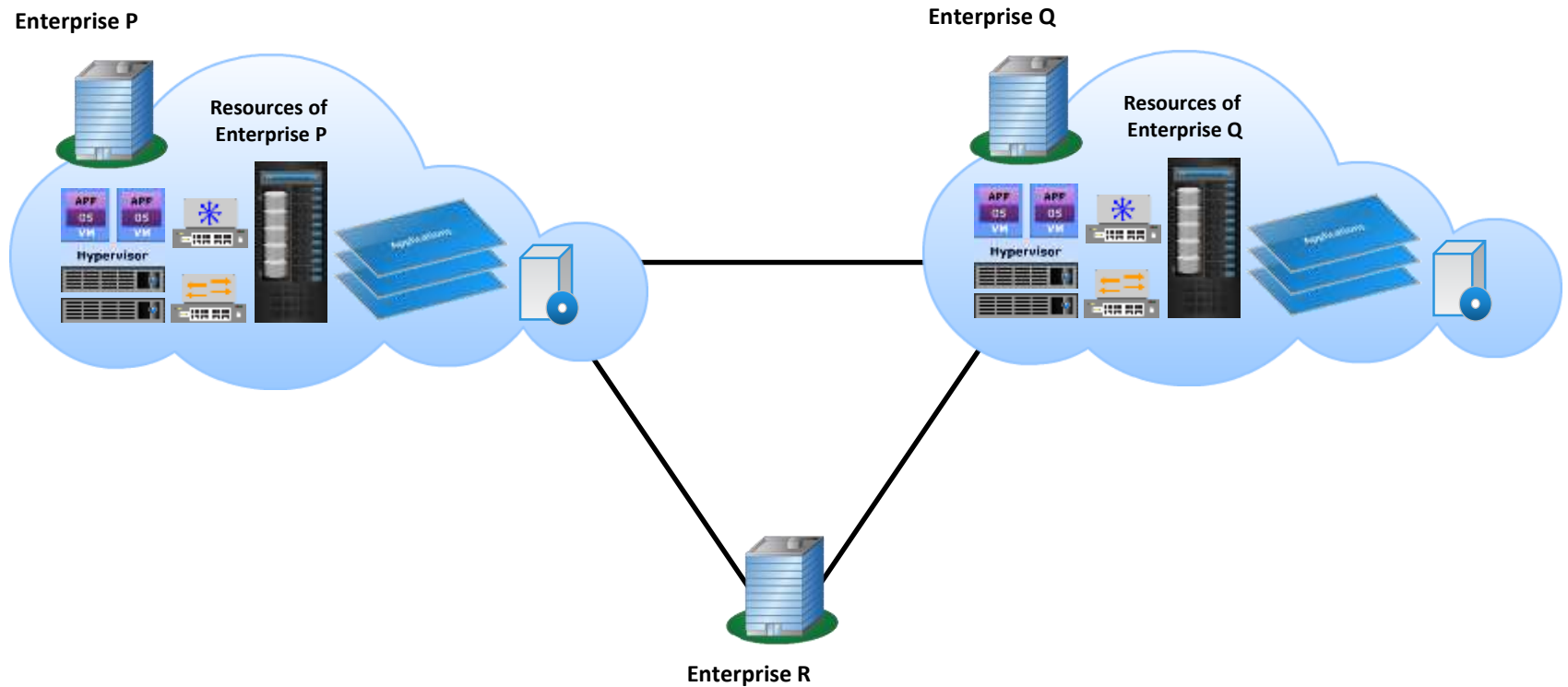


Private Cloud



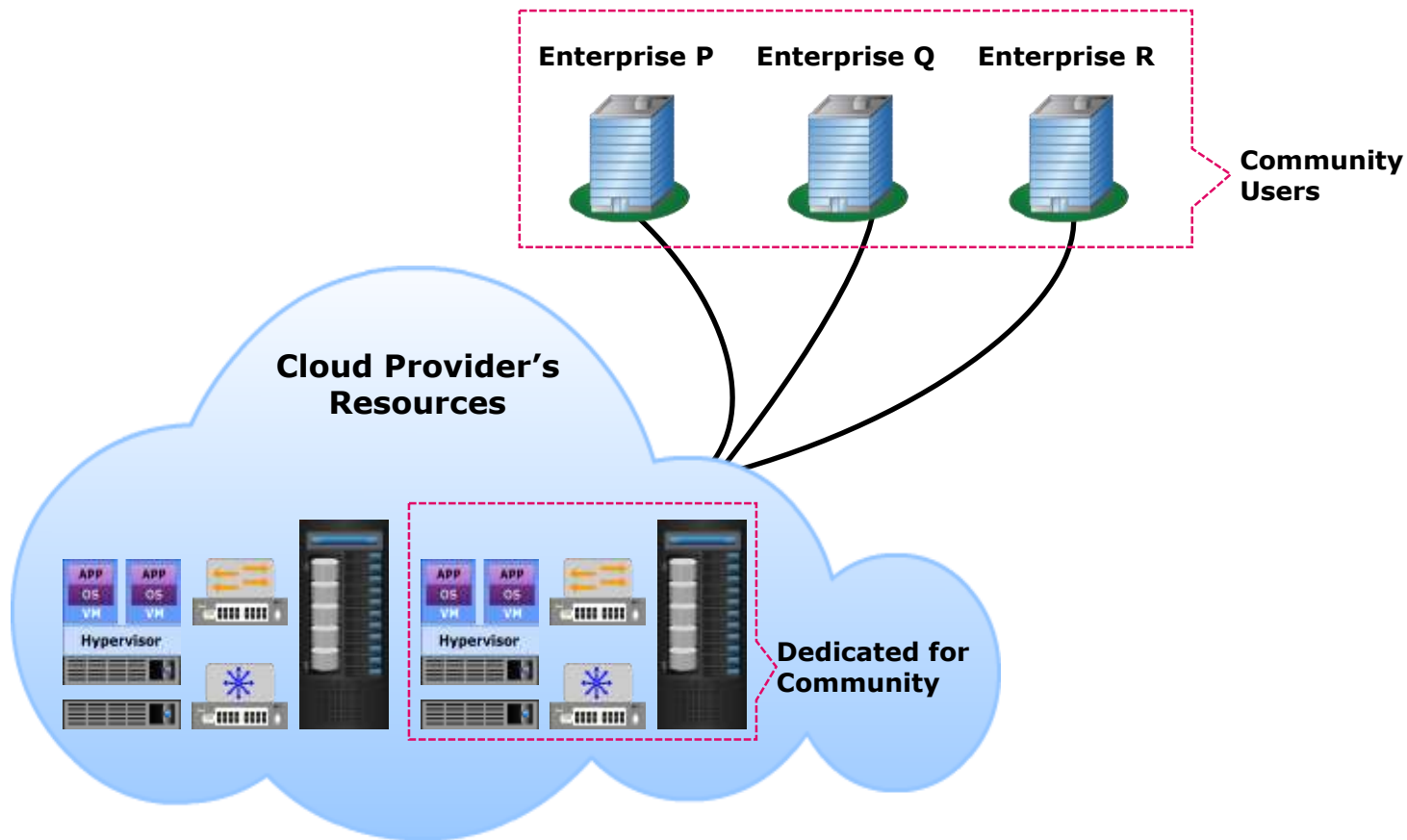
Community Cloud

- On-premise Community Cloud

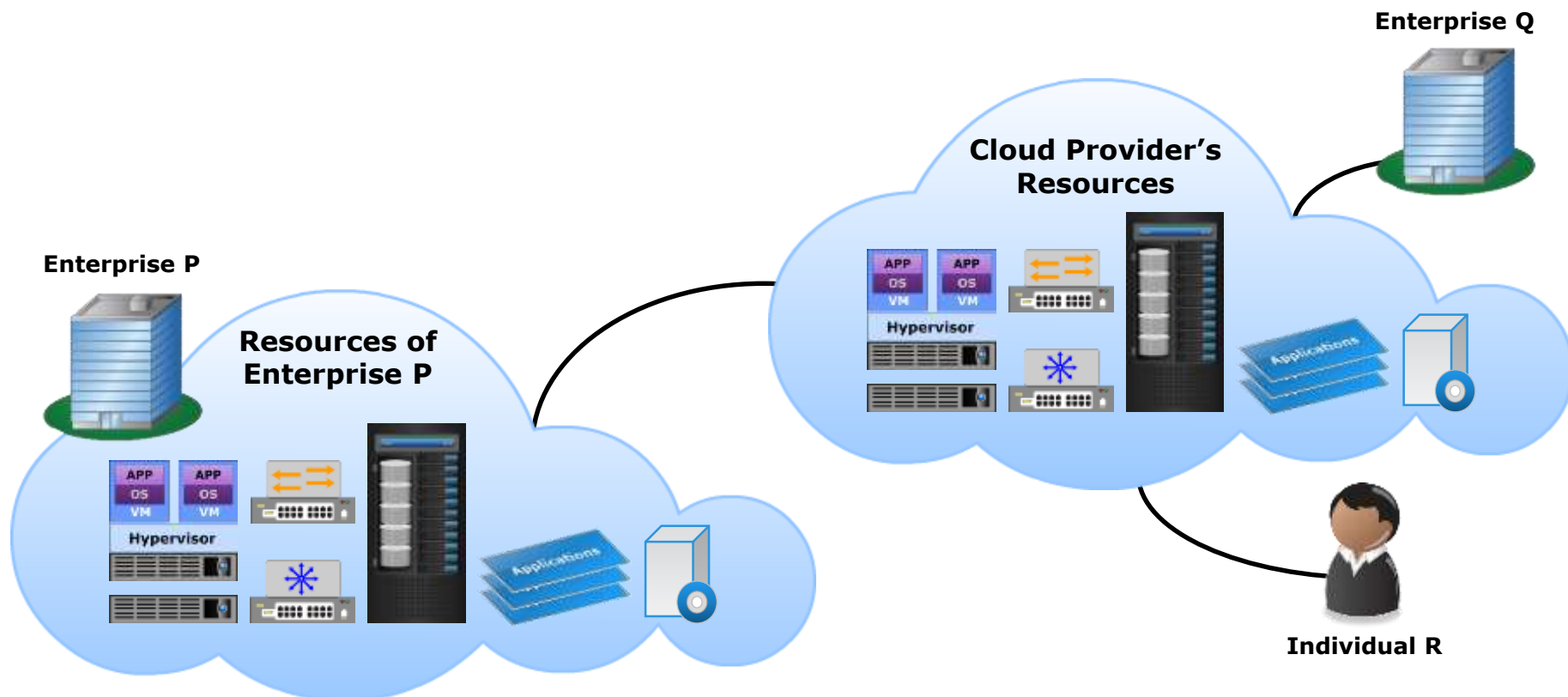


Community Cloud

- Externally-hosted Community Cloud



Hybrid Cloud



3

Industrial Solutions




Hadoop *hadoop*

- Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.
- It is a flexible and highly-available architecture for large scale computation and data processing on a network of commodity hardware.
- Designed to answer the question: “**How to process big data with reasonable cost and time?**”

Origin of Hadoop (1)



- Search Engine in 1990's


MetaCrawler Parallel Web Search Service
 by [Erik Selberg](#) and [Oren Etzioni](#)

Try the new [MetaCrawler Beta!](#)
 If you're searching for a person's home page, try [Abol!](#)

[Examples](#) • [Beta Site](#) • [Add Site](#) • [About](#)

Search for:

☐ as a Phrase ☒ All of these words ☐ Any of these words

For better results, please specify:

Search Region: Search Sites:

Performance parameters:

Max wait: minutes Match type:

[About](#) | [Help](#) | [Problem](#) | [Add Site](#) | [Search](#)
webmaster@metacrawler.com
© Copyright 1995, 1996 Erik Selberg and Oren Etzioni


 search reviews city.net **new!** live! reference?

[excite home](#) [maps](#) [news](#) [people finder](#)

Excite Search: twice the power of the competition.

What:

Where:

Researching stocks?
 Buying a car?
 Planning a wedding?
[Check out ExciteSearcng Tours.](#)

[Bill Mitchell](#)
 Satire that clicks!

Excite Reviews: site reviews by the web's best editorial team.

• Arts	• Entertainment	• Money	• Regional
• Business	• Health	• News & Reference	• Science
• Computing	• Hobbies	• Personal Pages	• Shopping
• Education	• Life & Style	• Politics & Law	• Sports

[Serious Sports Fans Only \\$1,000,000 in Cash and Prizes!](#)
[For serious sports fans only! Play Fantasy Football!](#)


It's amazing where Go Get It will get you.

Find:

[Enhance your search.](#)





[New Search](#) • [Top News](#) • [Sites by Subject](#) • [Top 5% Sites](#) • [City Guide](#) • [Pictures & Sounds](#)
[PeopleFind](#) • [Point Review](#) • [Road Maps](#) • [Software](#) • [About Lycos](#) • [Club Lycos](#) • [Help](#)

[Add Your Site to Lycos](#)

Copyright © 1996 Lycos™, Inc. All Rights Reserved.
 Lycos is a trademark of Carnegie Mellon University.
[Questions & Comments](#)

[HELP](#) [WIRED NEWS](#) [HOTWIRED](#) [WIRED MAGAZINE](#) [SUCK.COM](#)


The WIRED Search Center

from by:

For more options use [Advanced Search](#)

Date:
 Country:

☐ Images ☐ Audio ☐ Video ☐ Combinations

Return Results:

[Find](#) [Advanced Search](#) [People](#) [Contact Advertisers](#)

[Sandbox With Entertainment](#)
[Shop WIRED Holiday Gift Guide](#)
SOMETHING HAS SURVIVED.
[Find more details](#)

[Trip](#)
[Cybernetic Outpost](#)
[Microsoft Expedia Travel](#)
[ON SALE](#)

Origin of Hadoop (2)



- Search Engine in 1998 and 2010's



Origin of Hadoop (3)



2005: Doug Cutting and Michael J. Cafarella developed Hadoop to support distribution for the [Nutch](#) search engine project.



The project was funded by Yahoo.

2006: Yahoo gave the project to Apache Software Foundation.

Origin of Hadoop (4)



2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*



2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.



2006

Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach,
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

{fay,jeff,sanjay,wilson,deborah,mike,tushar,andy,gruber}@google.com

Google, Inc.

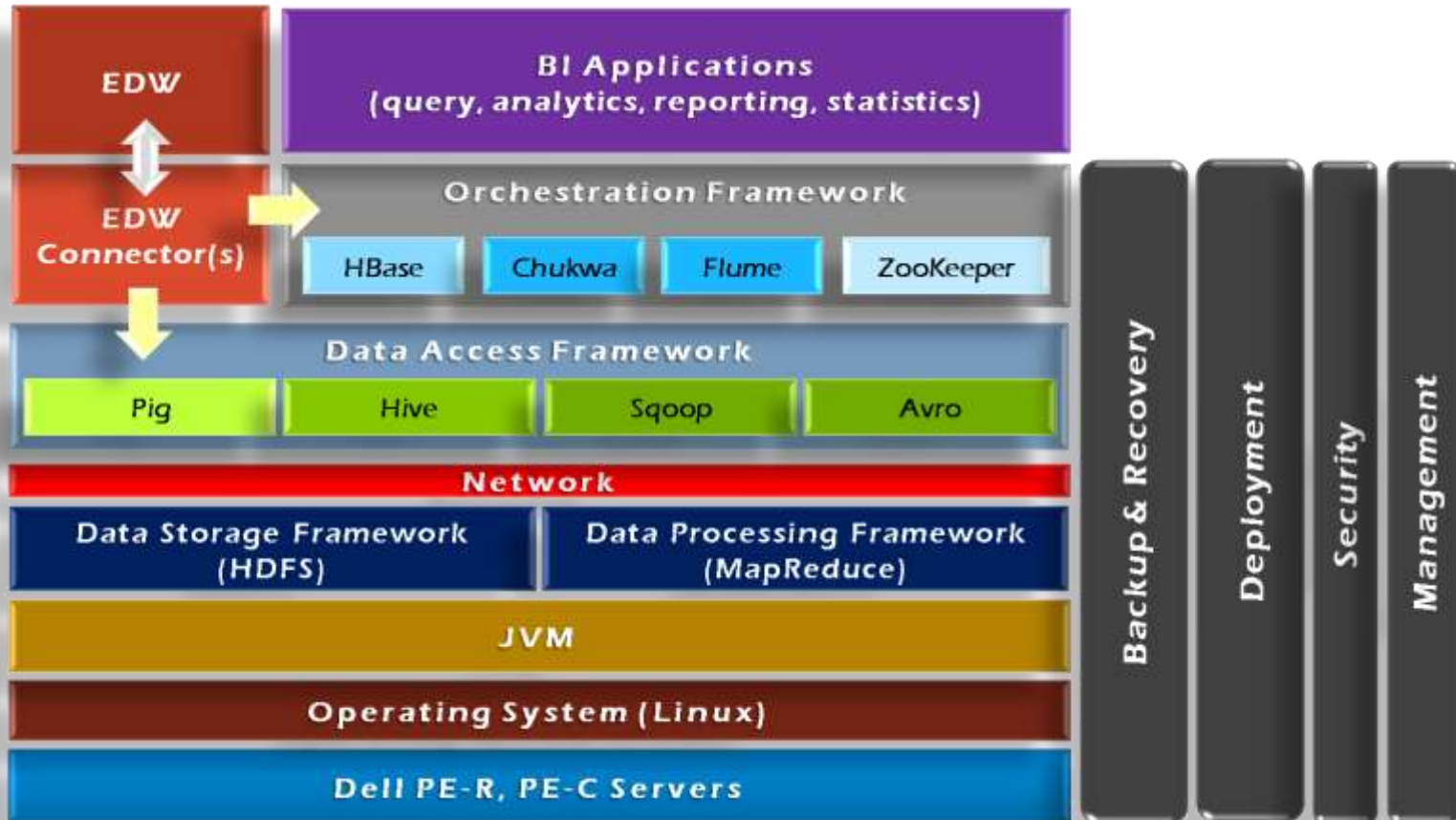


Abstract

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large number of nodes and to support a very large number of users. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Earth. These applications place very different demands on Bigtable, both in terms of data size (from URLs to

achieved scalability and high performance, but Bigtable provides a different interface than such systems. Bigtable does not support a full relational data model; instead, it provides clients with a simple data model that supports dynamic control over data layout and format, allows clients to reason about the locality properties of data represented in the underlying storage. Data is indexed using row and column names that can be arbitrary strings. Bigtable also treats data as uninterpreted bit

Hadoop Framework



Google Google Cloud Platform



Compute Google Cloud Platform



Compute Engine: Virtual machines hosted on Google's infrastructure - [Infrastructure-as-a-Service](#)



App Engine: Deploy your code directly to a fully-managed platform - [Platform-as-a-Service](#)



Container Engine: Run Docker container cluster on Google Cloud Platform – [Container-as-a-Service](#)

Storage Google Cloud Platform



Cloud SQL: Full SQL support for an online transaction processing (OLTP) system



Cloud Datastore: Store highly structured objects and query with SQL-like statements



Cloud Storage: Store immutable blobs larger than 10 MB, such as large images or videos



Cloud BigTable: High-performance, extremely scalable NoSQL database, scales to billions of entries

Amazon



- AWS is Amazon's umbrella description of all of their web-based technology services.
- Mainly infrastructure services:
 - Amazon Elastic Compute Cloud (EC2)
 - Amazon Simple Storage Service (S3)
 - Amazon Simple Queue Service (SQS)
 - Amazon CloudFront
 - Amazon SimpleDB

Amazon



Amazon



Database

DynamoDB

Predictable and Scalable NoSQL Data Store

ElastiCache

In-Memory Cache

RDS

Managed Relational Database

Redshift

Managed Petabyte-Scale Data Warehouse

Storage & CDN

S3

Scalable Storage in the Cloud

EBS

Networked Attached Block Device

CloudFront

Global Content Delivery Network

Glacier

Archive Storage in the Cloud

Storage Gateway

Integrates On-Premises IT with Cloud Storage

Import Export

Ship Large Datasets

Cross-Service

Support

Phone & email fast-response 24X7 Support

Marketplace

Buy and sell Software and Apps

Management Console

UI to manage AWS services

SDKs, IDE kits and CLIs

Develop, integrate and manage services

Analytics

Elastic MapReduce

Managed Hadoop Framework

Kinesis

Real-Time Data Stream Processing

Data Pipeline

Orchestration for Data-Driven Workflows

Compute & Networking

EC2

Virtual Servers in the Cloud

VPC

Virtual Secure Network

ELB

Load balancing Service

WorkSpaces

Virtual Desktops in the cloud

Auto Scaling

Automatically scale up and down

DirectConnect

Dedicated Network Connection to AWS

Route 53

Scalable Domain Name System

Deployment & Management

CloudFormation

Templated AWS Resource Creation

CloudWatch

Resource and Application Monitoring

Elastic Beanstalk

AWS Application Container

IAM

Secure AWS Access Control

CloudTrail

User Activity Logging

OpsWorks

DevOps Application Management Service

CloudHSM

Hardware-based key storage for compliance

App Services

CloudSearch

Managed Search Service

Elastic Transcoder

Easy-to-use Scalable Media Transcoding

SES

Email Sending Service

SNS

Push Notification Service

SQS

Message Queue Service

SWF

Workflow Service for Coordinating App Components

AppStream

Low-latency Application Streaming

AWS Global Physical Infrastructure
(Geographical Regions, Availability Zones, Edge Locations)

AWS Management Console



Amazon S3

Amazon EC2

Amazon VPC

Amazon Elastic MapReduce

Amazon CloudFront

Amazon RDS

Amazon SNS

Navigation

Region: US East

EC2 Dashboard

INSTANCES

Instances

Spot Requests

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Snapshots

NETWORKING & SECURITY

Elastic IPs

Security Groups

Placement Groups

Load Balancers

Key Pairs

My Instances

Launch Instance

Instance Actions

Reserved Instances

Show/Hide

Refresh

Help

Viewing:

All Instances

All Instance Types

	Name	Instance	Type	Status	Lifecycle	Public DNS
<input checked="" type="checkbox"/>	Web Server	i-841948e9	m1.small	running	normal	ec2-67-202-15-66.compute-1.

1 EC2 Instance selected

EC2 Instance: i-841948e9

Description

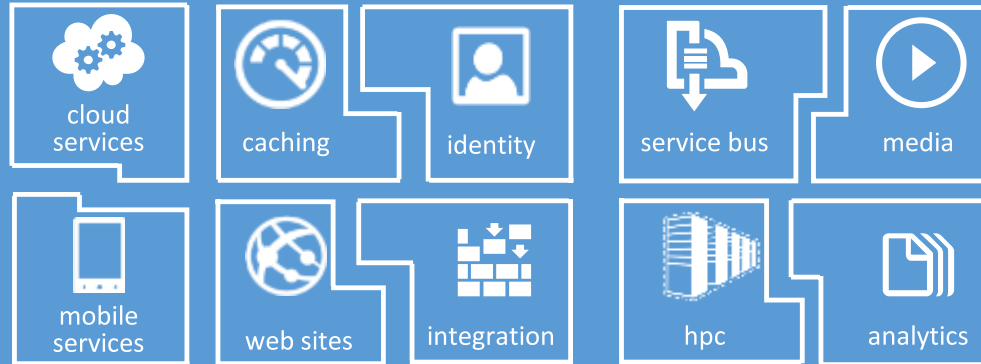
Monitoring

Tags

AMI ID:	ami-08728661	Zone:	us-east-1b
Security Groups:	80_22_open	Type:	m1.small
Status:	running	Owner:	043708602122
VPC ID:	-	Subnet ID:	-
Virtualization:	paravirtual	Placement Group:	
Reservation:	r-7de68517	RAM Disk ID:	-
Platform:	-	Key Pair Name:	GSG_Keypair
Kernel ID:	aki-407d9529	Monitoring:	basic
AMI Launch Index:	0	Elastic IP:	-
Root Device:	/dev/sda1	Root Device Type:	ebs

Microsoft Azure (1) Microsoft Azure

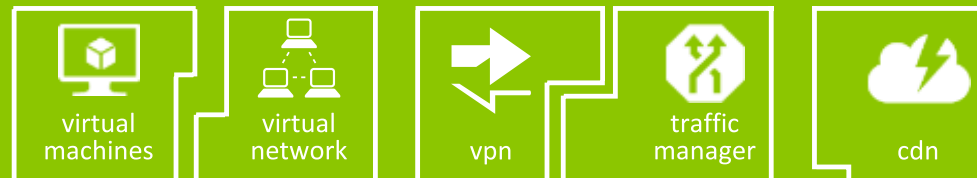
app services



data services



infrastructure services

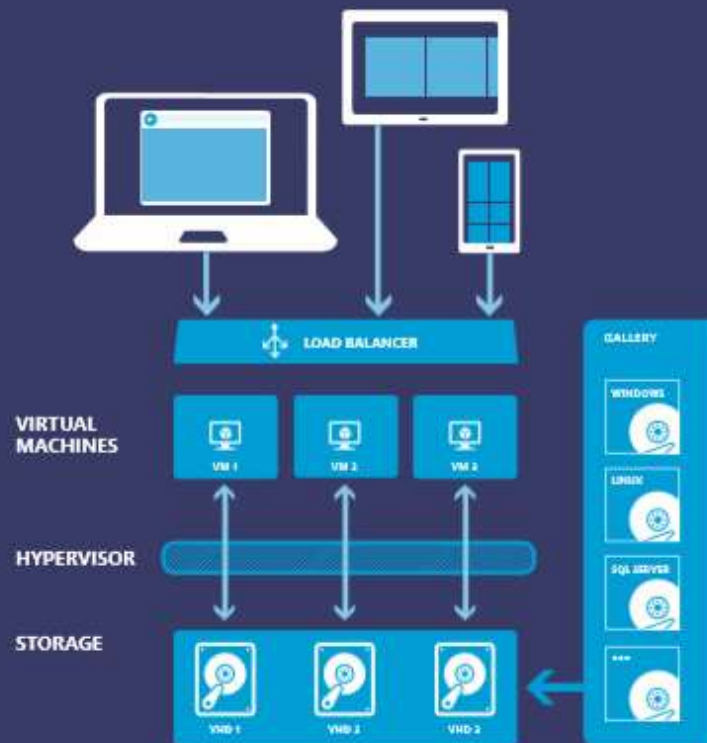


Microsoft Azure (2) Microsoft Azure



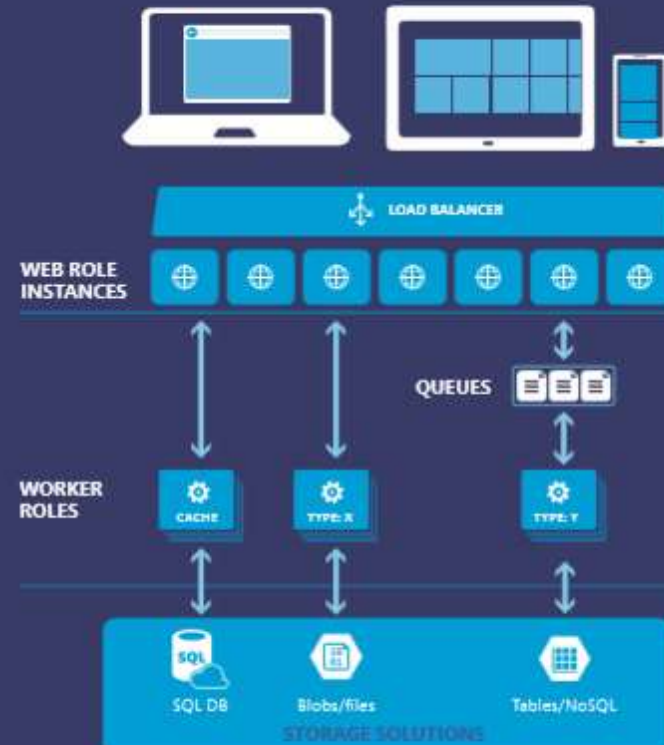
1. Virtual Machines

The basic cloud building block that gives you full access to a virtual machine with persistent storage that you completely own and control. You deploy, manage and architect resilience yourself across collections of VMs. These are most similar to VMs on-premise and are the easiest way to move existing workloads to the cloud.

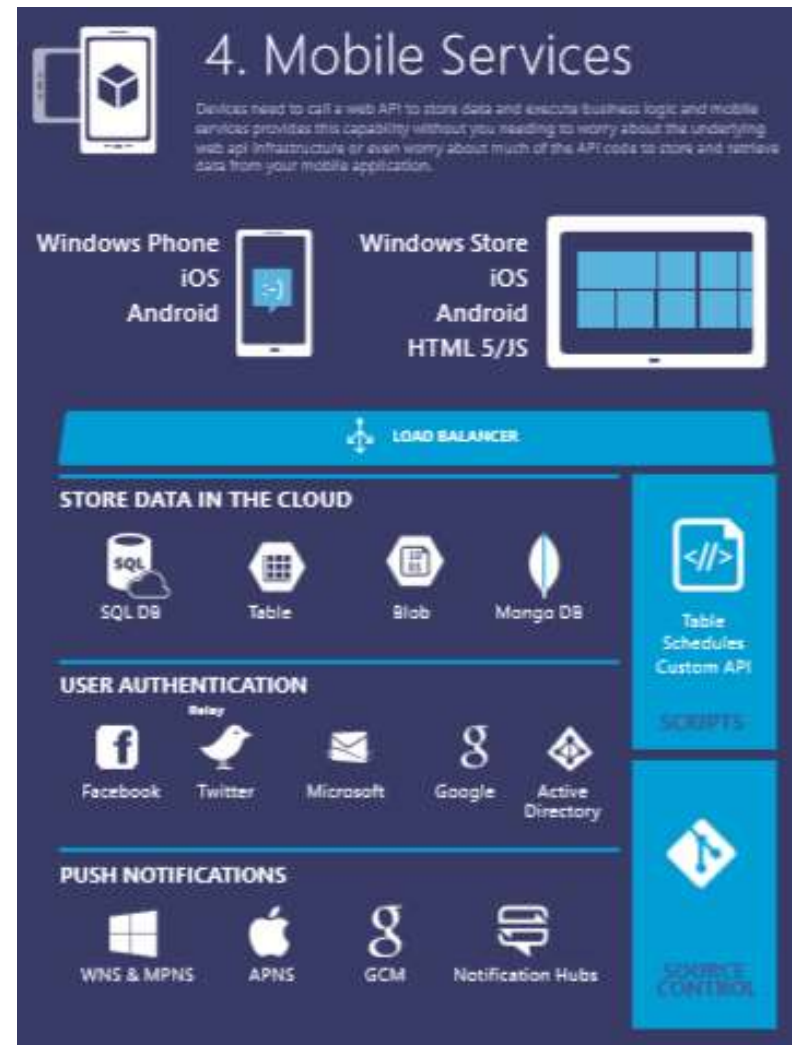
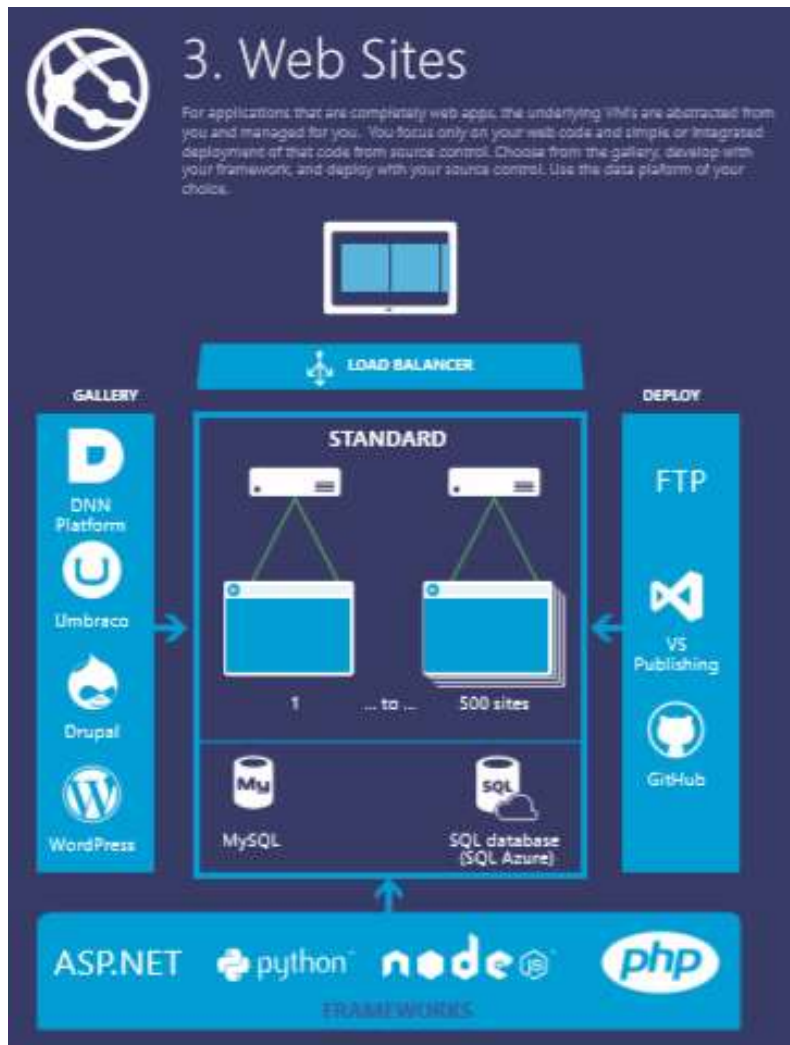


2. Cloud Services

Manage general purpose VMs that you have access to and can do quite low level configuration and deployment of additional software in the VM. VMs are stateless and you need to architect for and store state for your applications outside the VM. The Web Role is simply a worker role with its already installed/configured.



Microsoft Azure (3) Microsoft Azure



Aliyun Framework(1)



Support



Professional
Services



Training &
Certification



Cloud
Architects



Price Report



Enterprise
Email

Technical
Business Support



Analytics



ODPS



ADS



DPC



CDP



Application Middleware



OpenSearch



ACE



MTS



EDAS



MNS



ONS



PTS



SLS

Enterprise
Application



Access Control



RAM



Resource Management



CMS



CLI



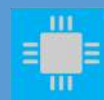
Security



Yundun

Platform
Services

Administration
& Security



Compute



ECS



ESS



SLB



Batch Compute



Storage



OSS



OAS



Database



RDS



DRDS



OTS



OCS



KVStore



Network



VPC



CDN

Core
Services



Global



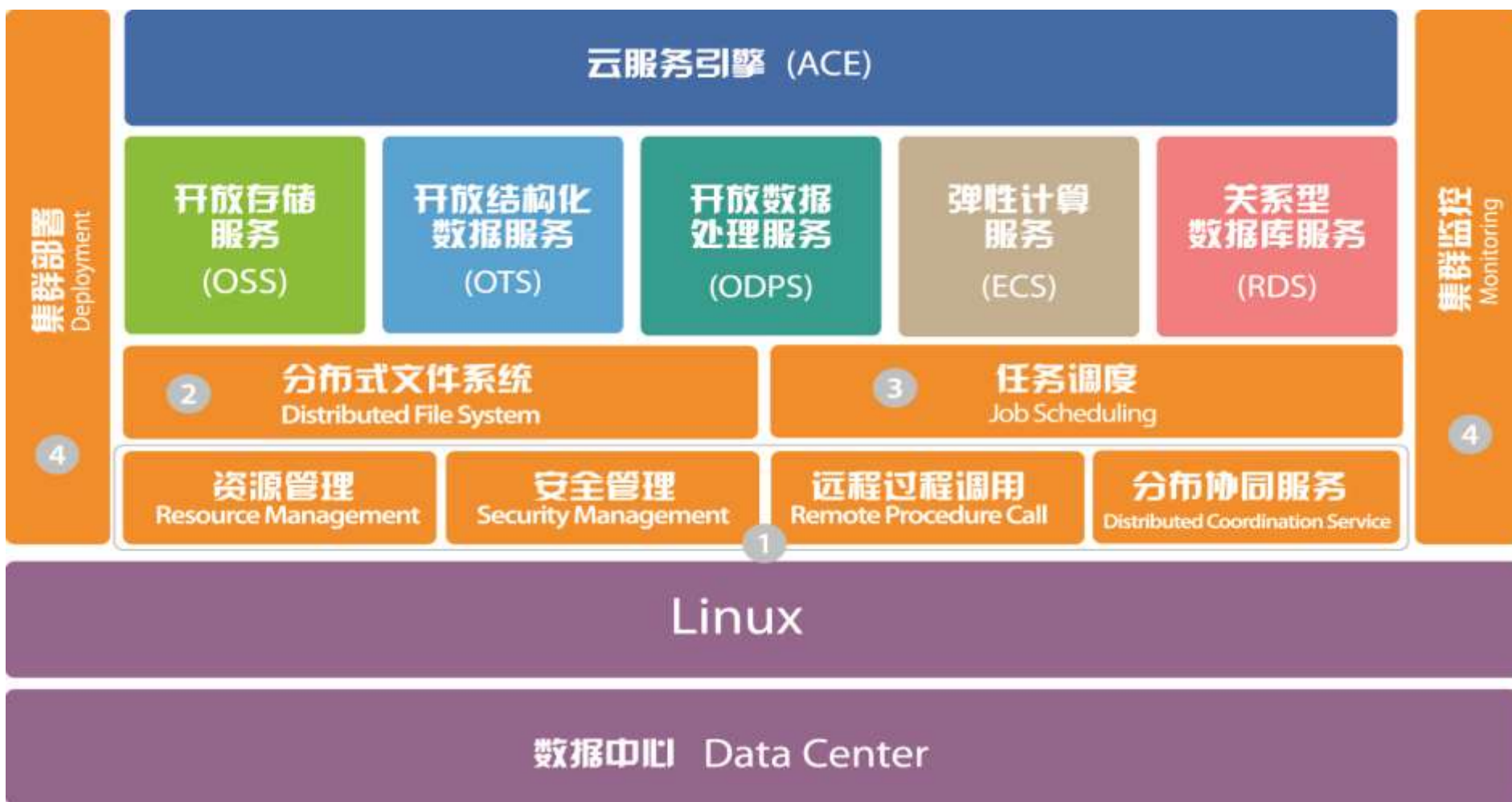
Regions



Availability
Zones

Infrastructure

Aliyun Framework (2)



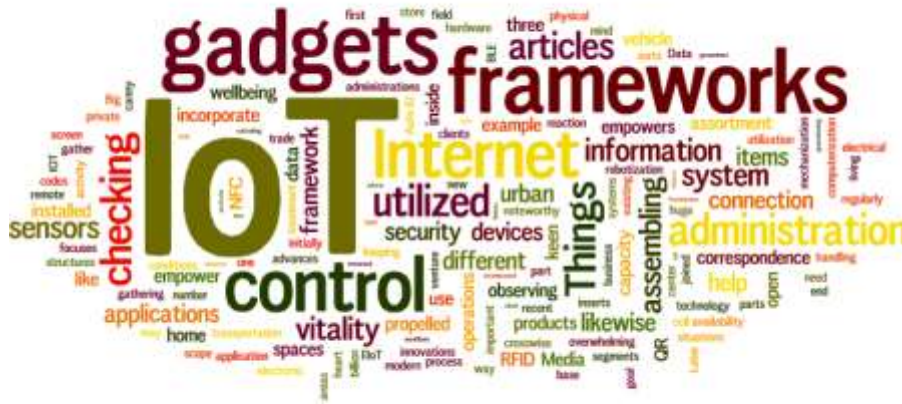
4

IoT



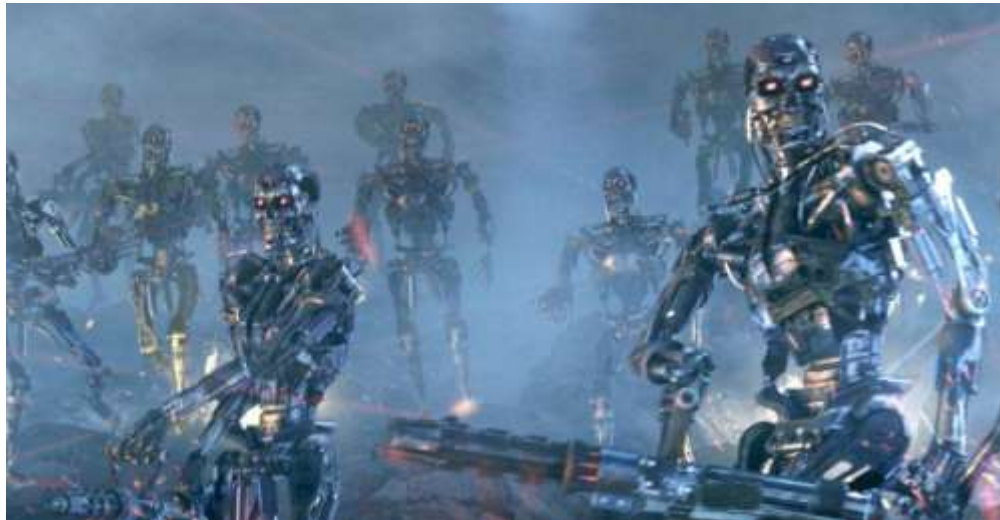
IoT (Internet of Things)

- The **Internet of Things (IoT)** is the network of physical objects—devices, vehicles, buildings and other items embedded with electronics, software, sensors, and network connectivity—that enables these objects to collect and exchange data.



Various names, One concept

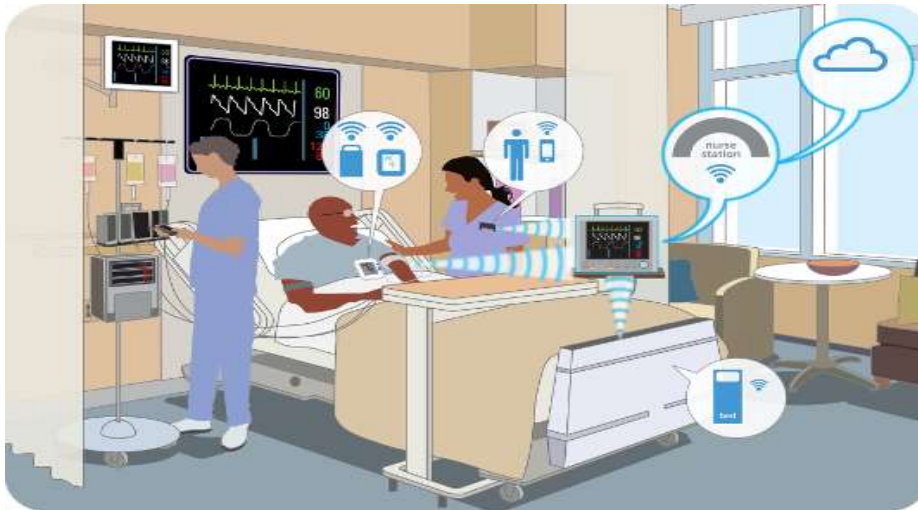
- M2M (Machine to Machine)
- “Internet of Everything” (Cisco Systems)
- “World Size Web” (Bruce Schneier)
- “Skynet” (Terminator movie)



Where is IoT



Wearable
Tech



Healthcare

Smart Appliances



IoT Access Many Industries



Healthcare and Life
Sciences



Municipal
Infrastructure



Smart Home



Retail



Manufacturing, Logistics &
Supply Chain



Agriculture

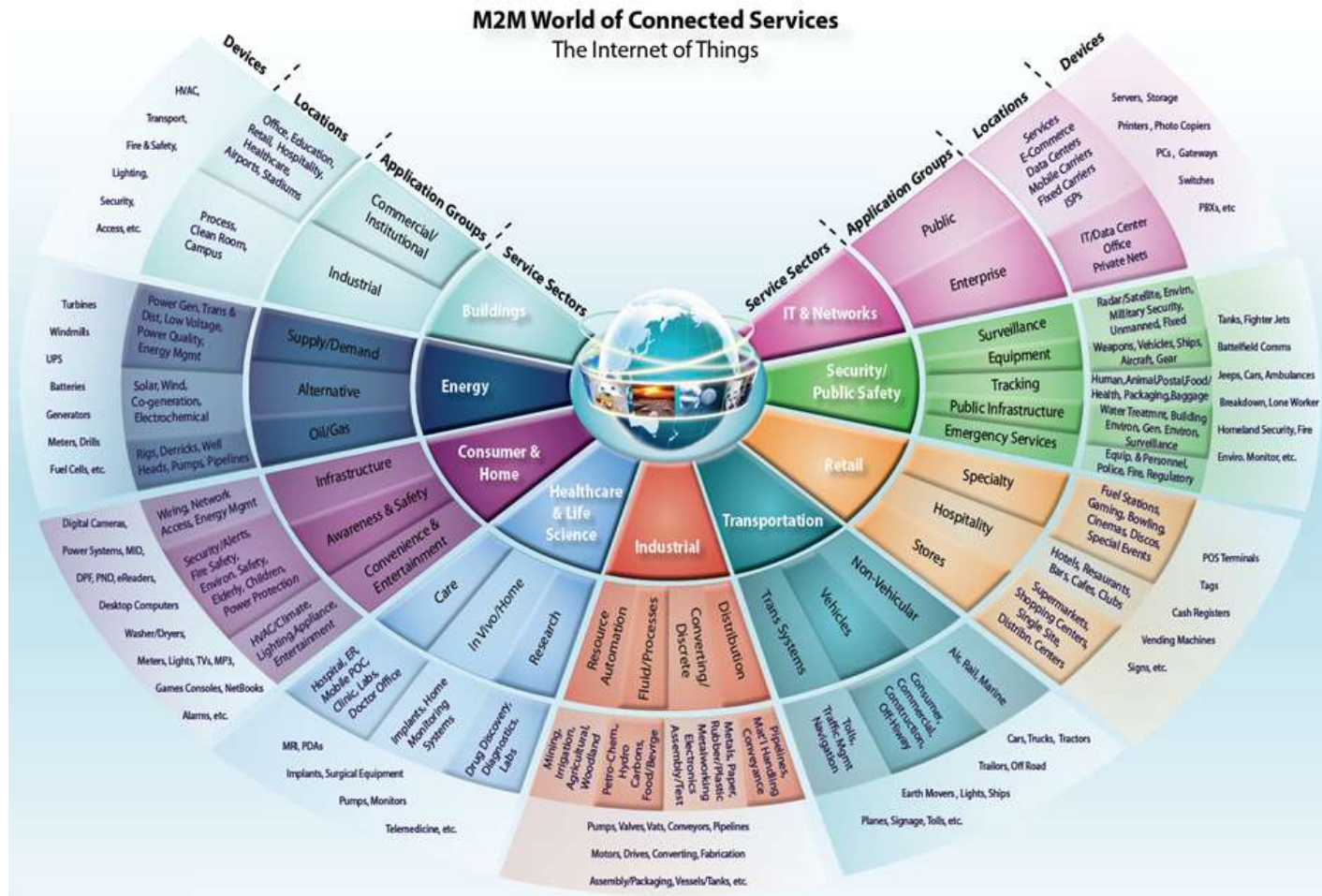


Education

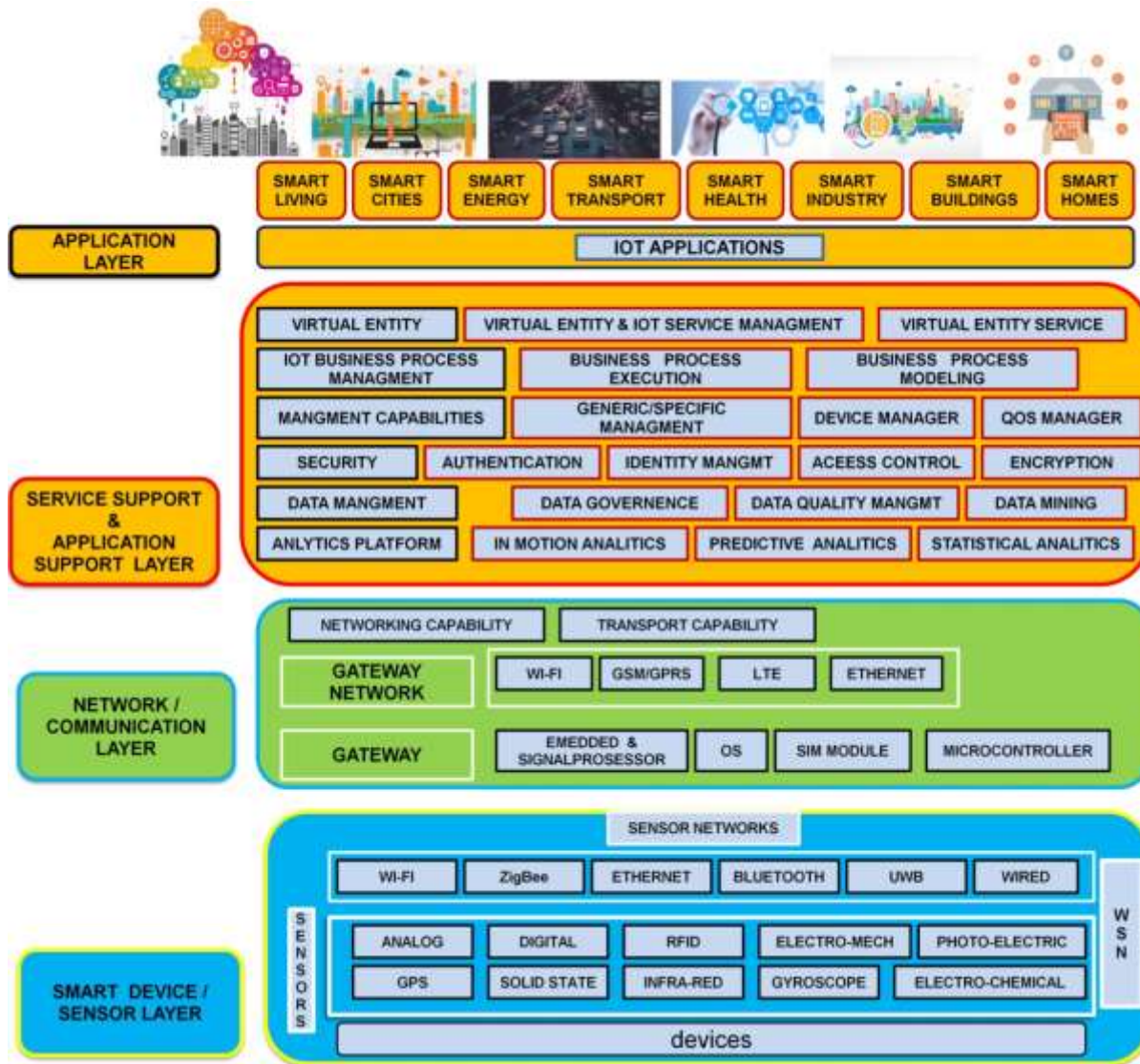


Automotive

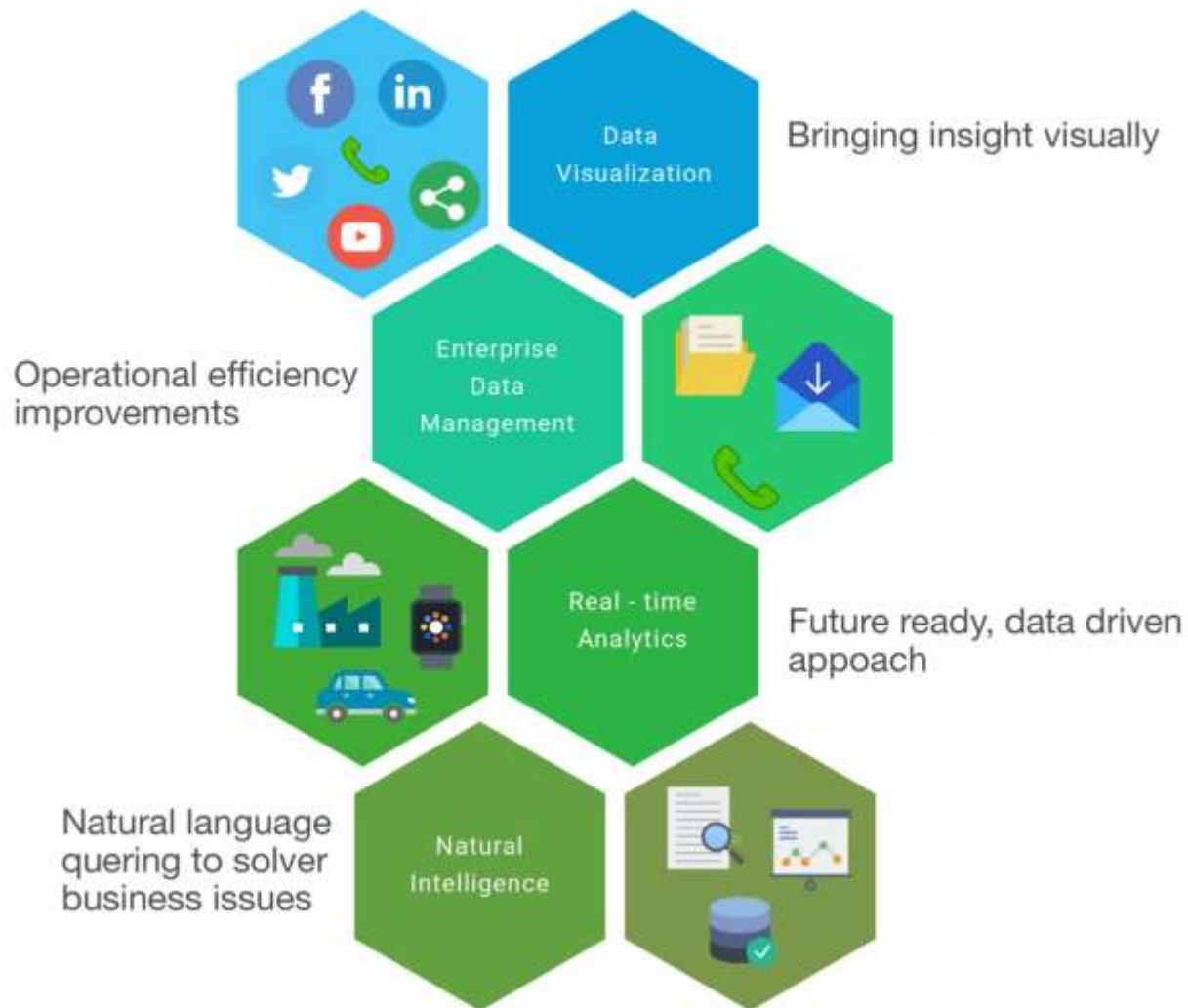
IoT Ecosystem



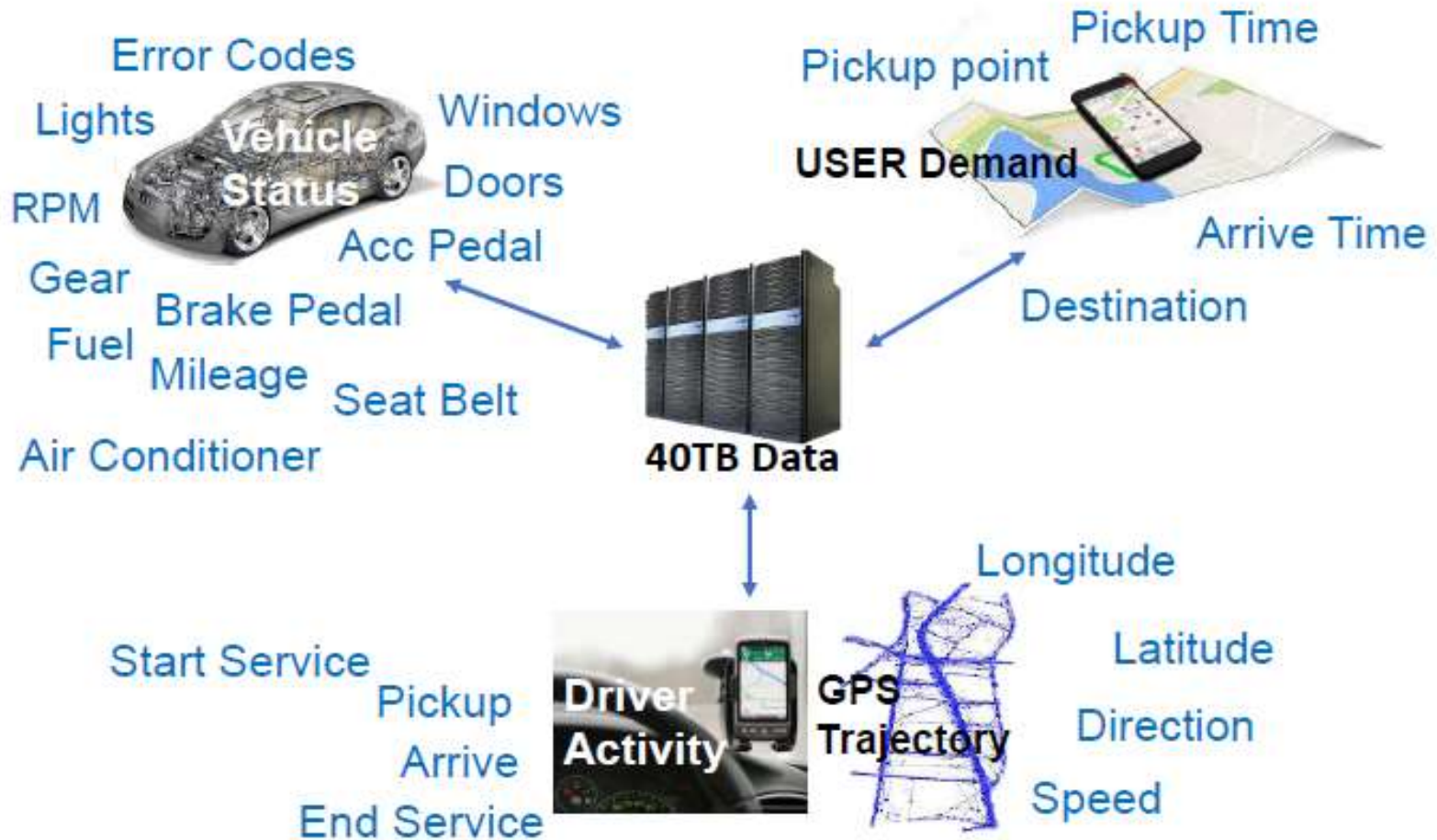
IoT Integration



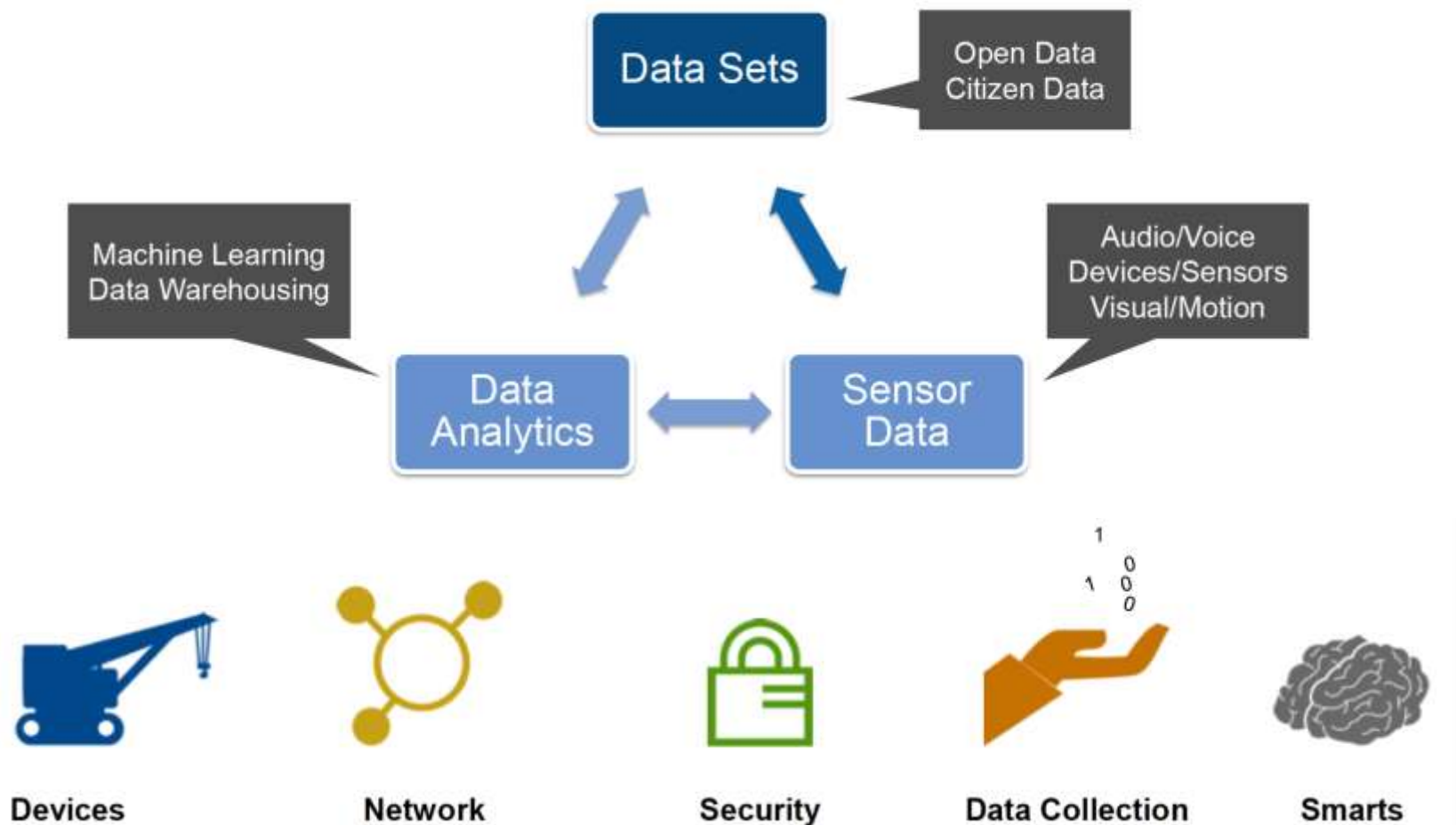
Big Data Problem in IoT



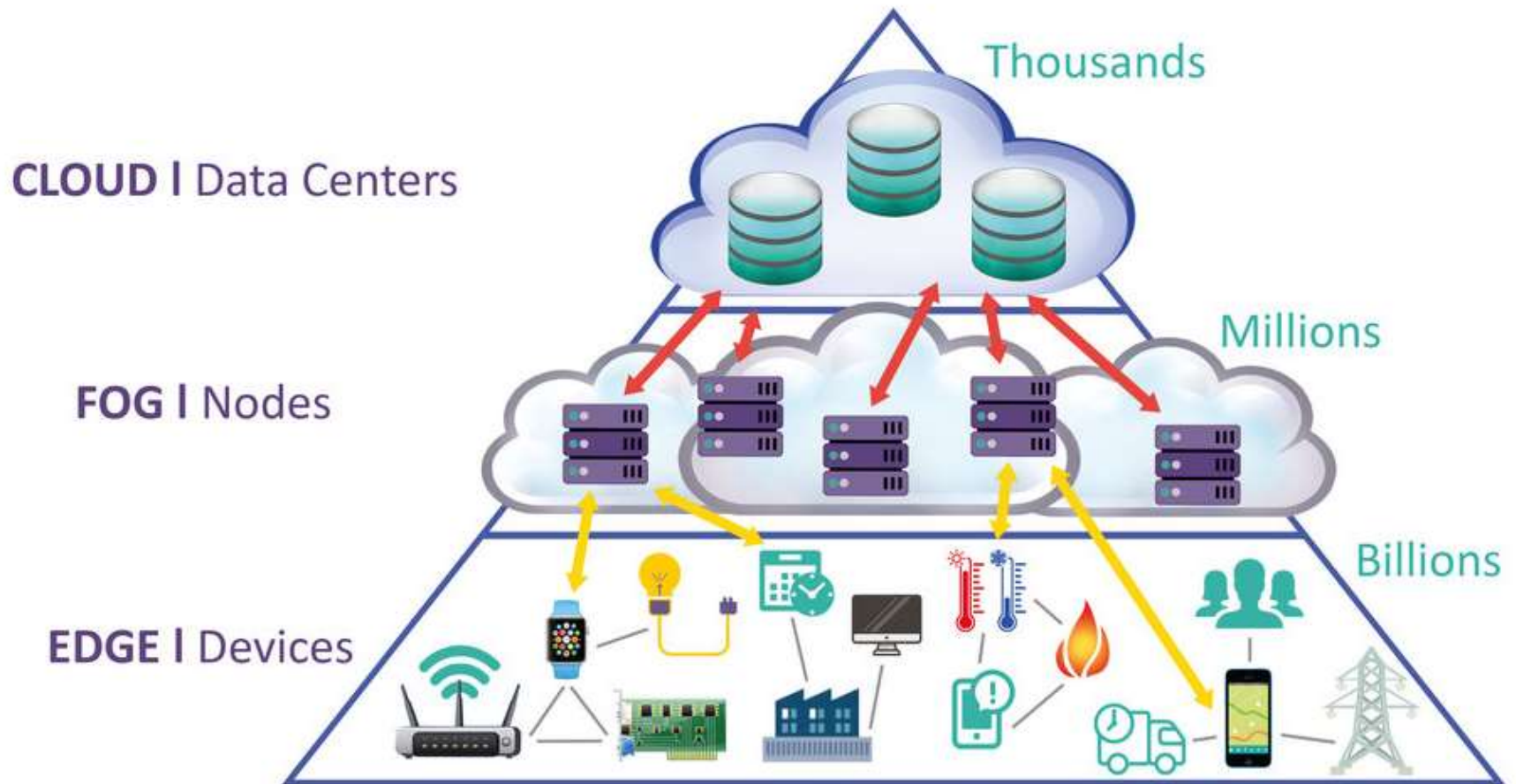
Big Data Problem in IoT (Example)



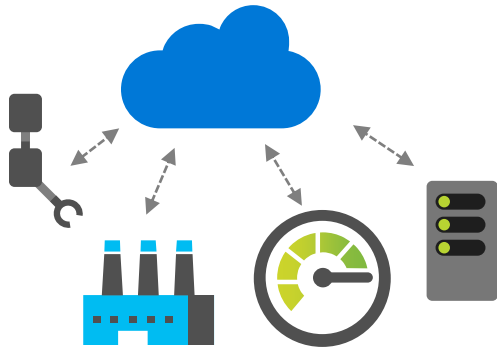
Big Data Processing in IoT



Cloud & Fog Fusion

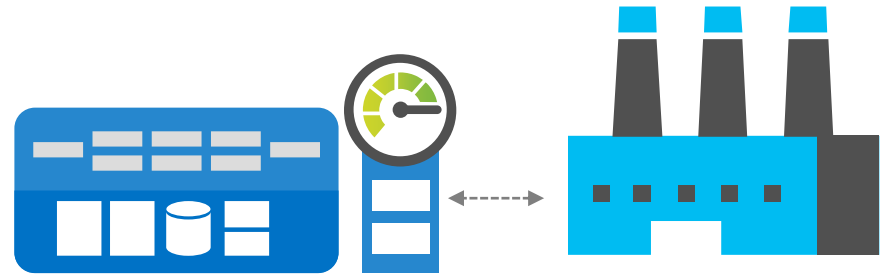


IoT in the cloud and on the edge



IoT in the Cloud

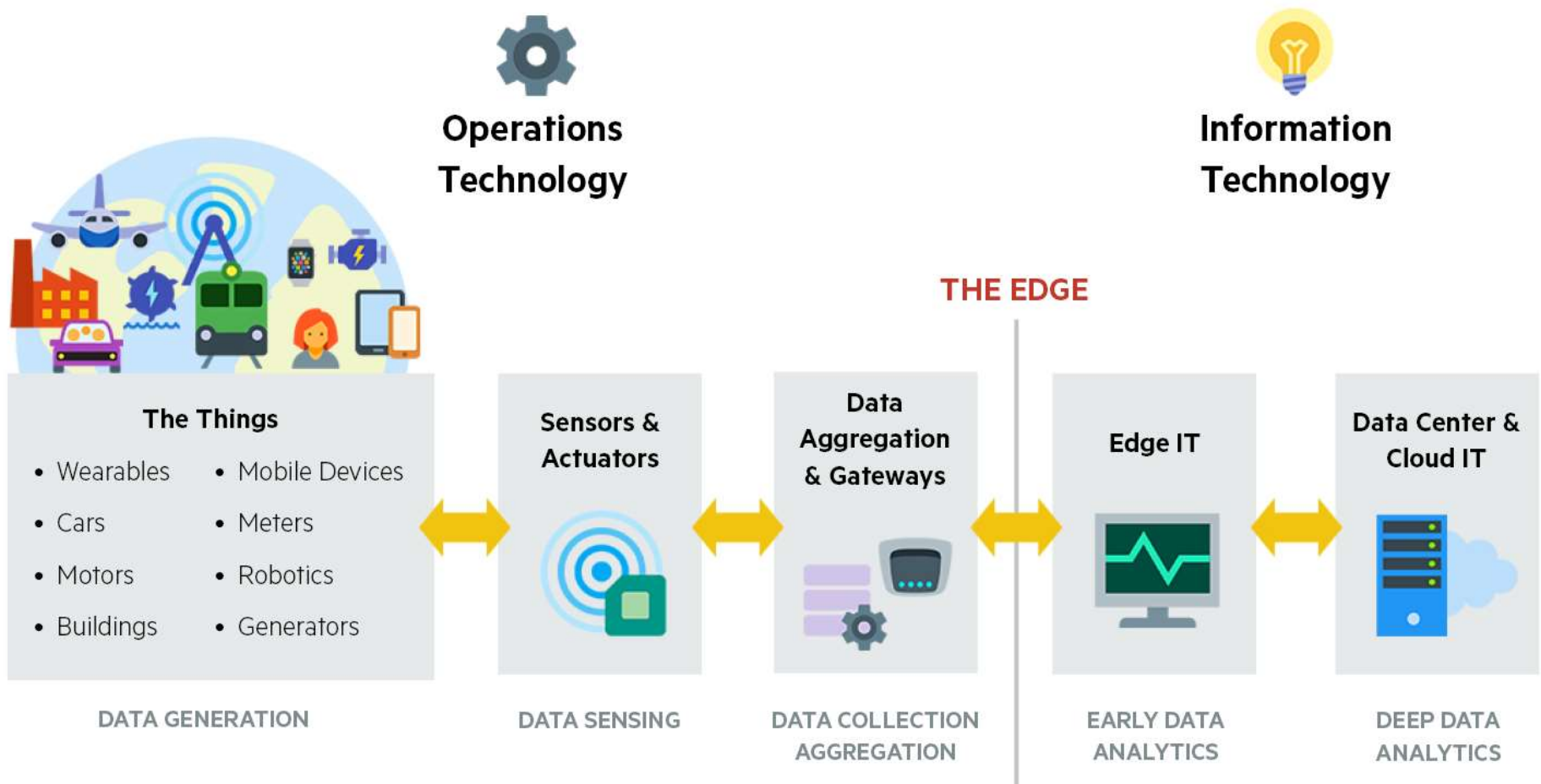
- Remote monitoring and control
- Merging remote data from across multiple IoT devices
- Near infinite compute and storage to train machine learning and other advanced AI tools



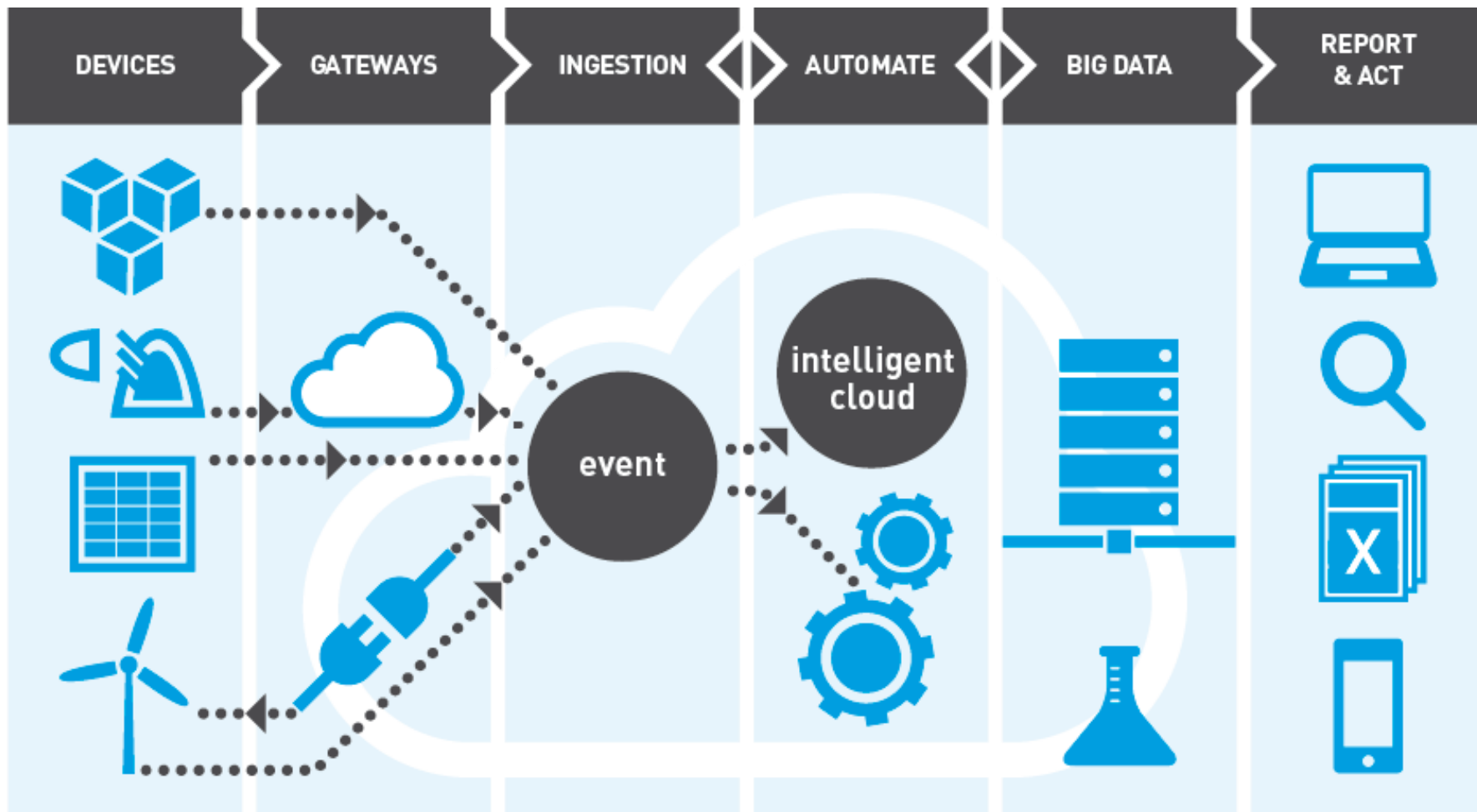
IoT on the Edge

- Low latency tight control loops require near real-time response
- Public internet inherently unpredictable
- Privacy of data and protection of IP

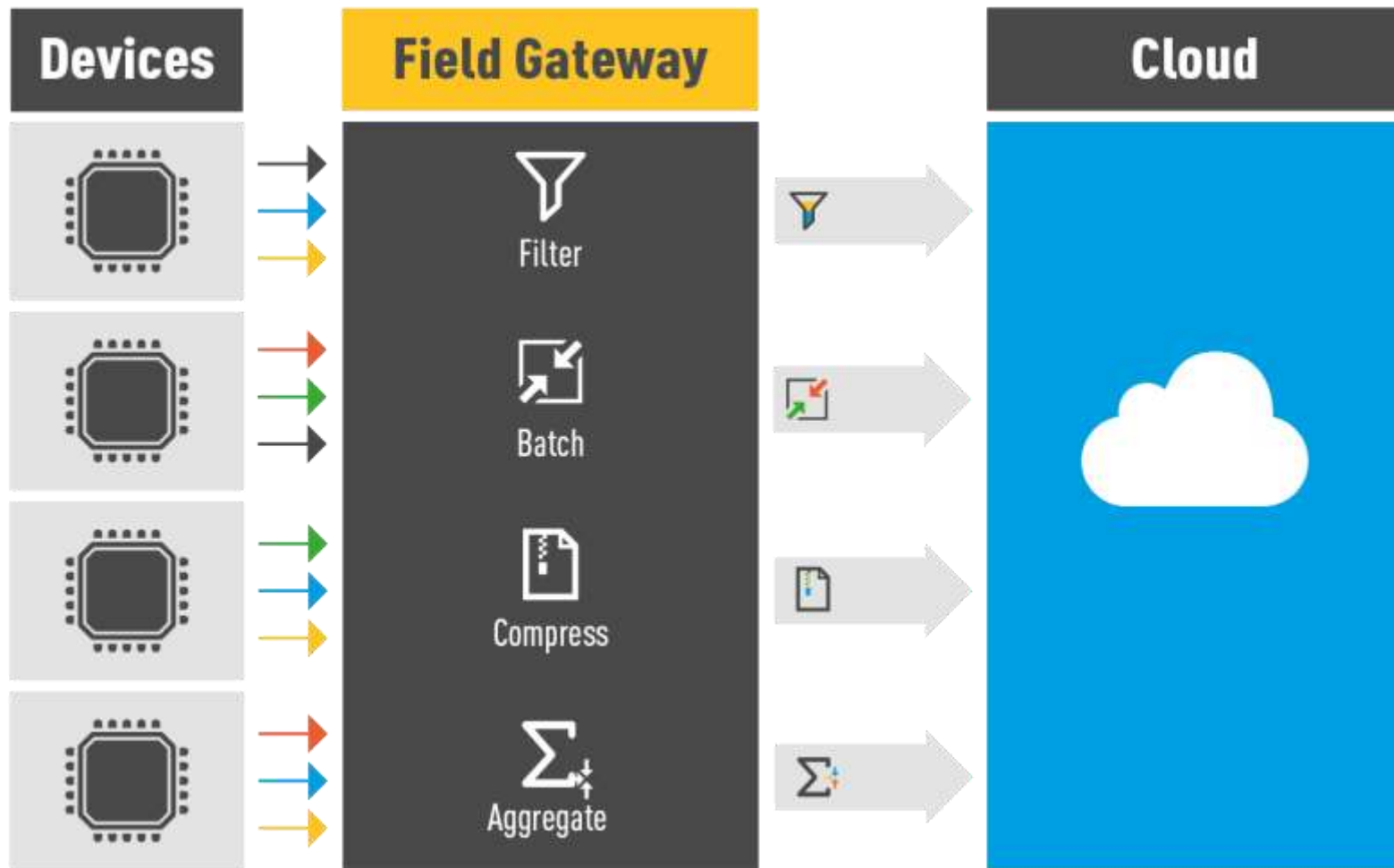
Fog/Edge Computing is the Primary Choice to Handle Real Time Data



IoT End-to-End Value Chain



Smart Gateway



Docker container

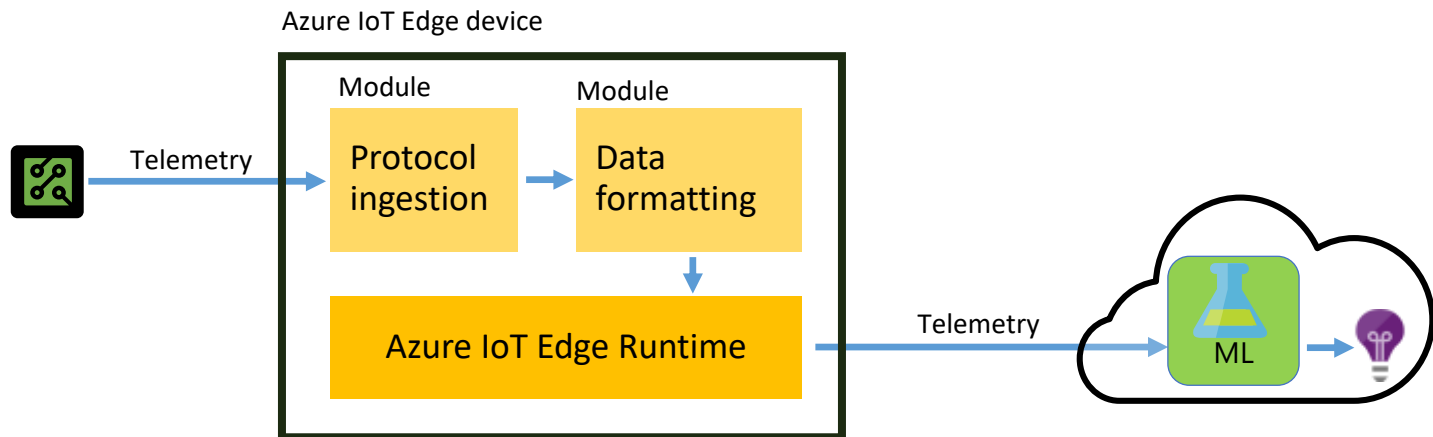
Edge Runtime manages modules

Modules add capabilities to the runtime

Each module performs an action

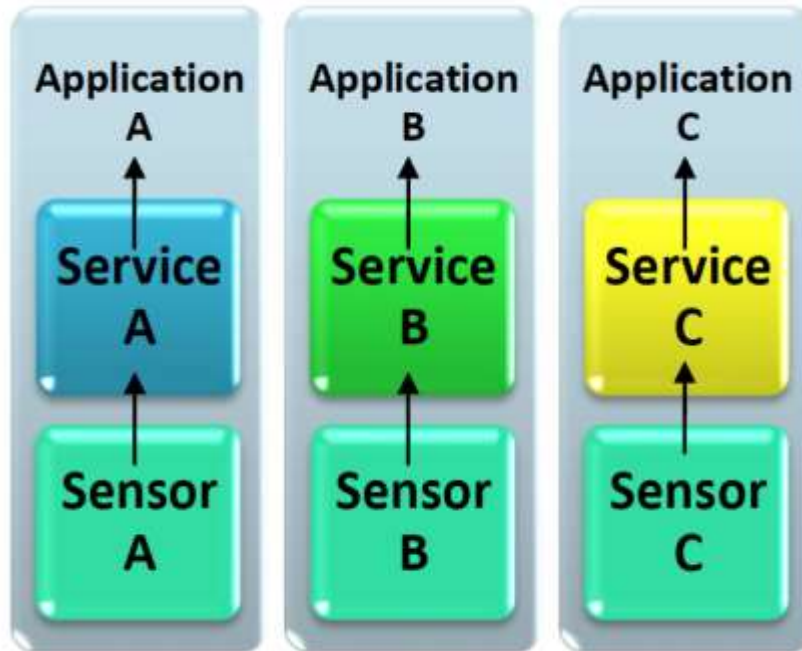
Chain of modules can be thought of as a data processing pipeline, solving an end to end scenario

Modules are Docker containers

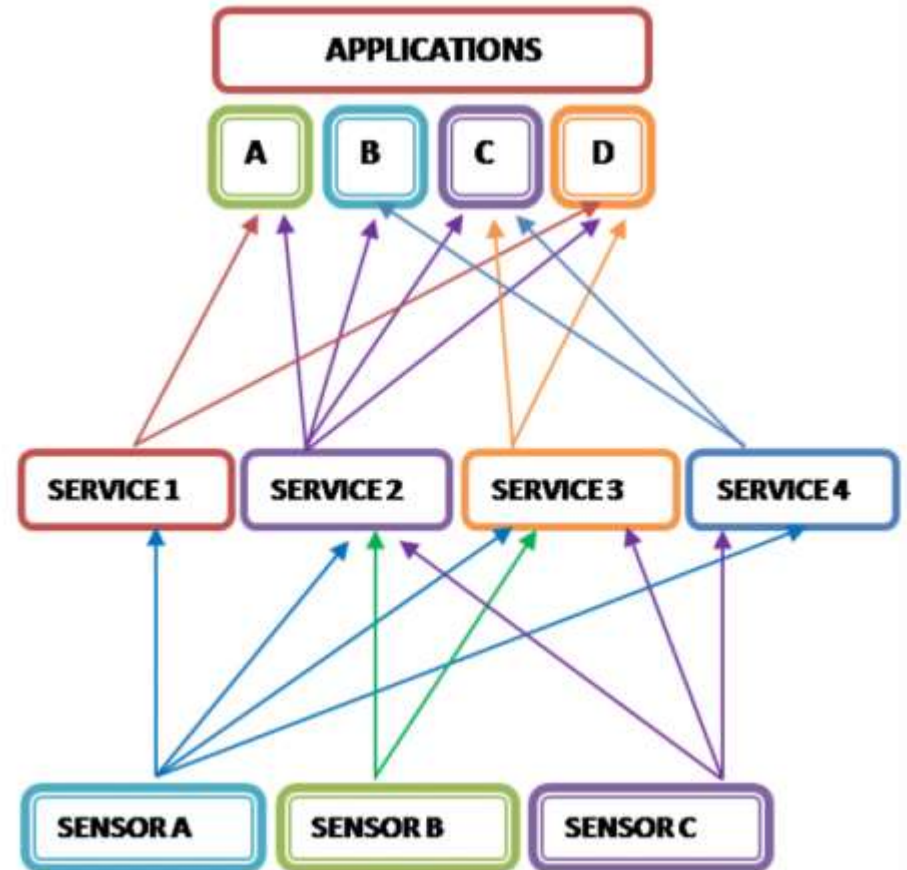


Interoperability problem in IoT

No Interoperation

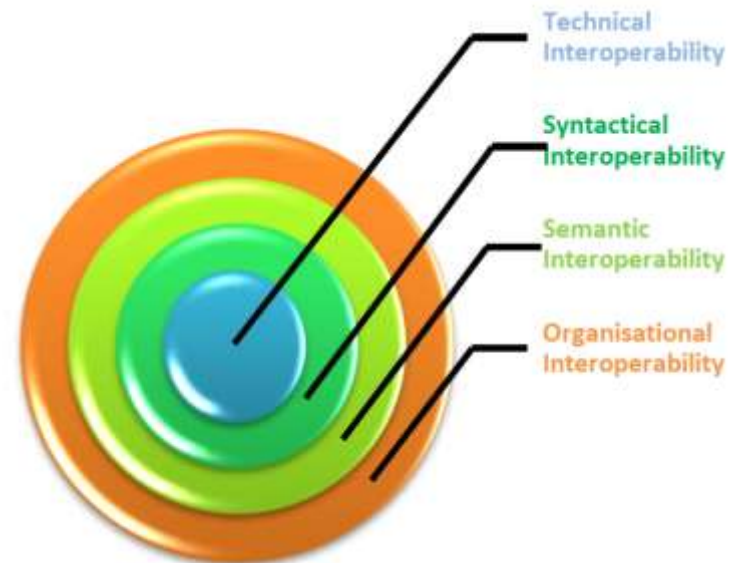


Interoperation

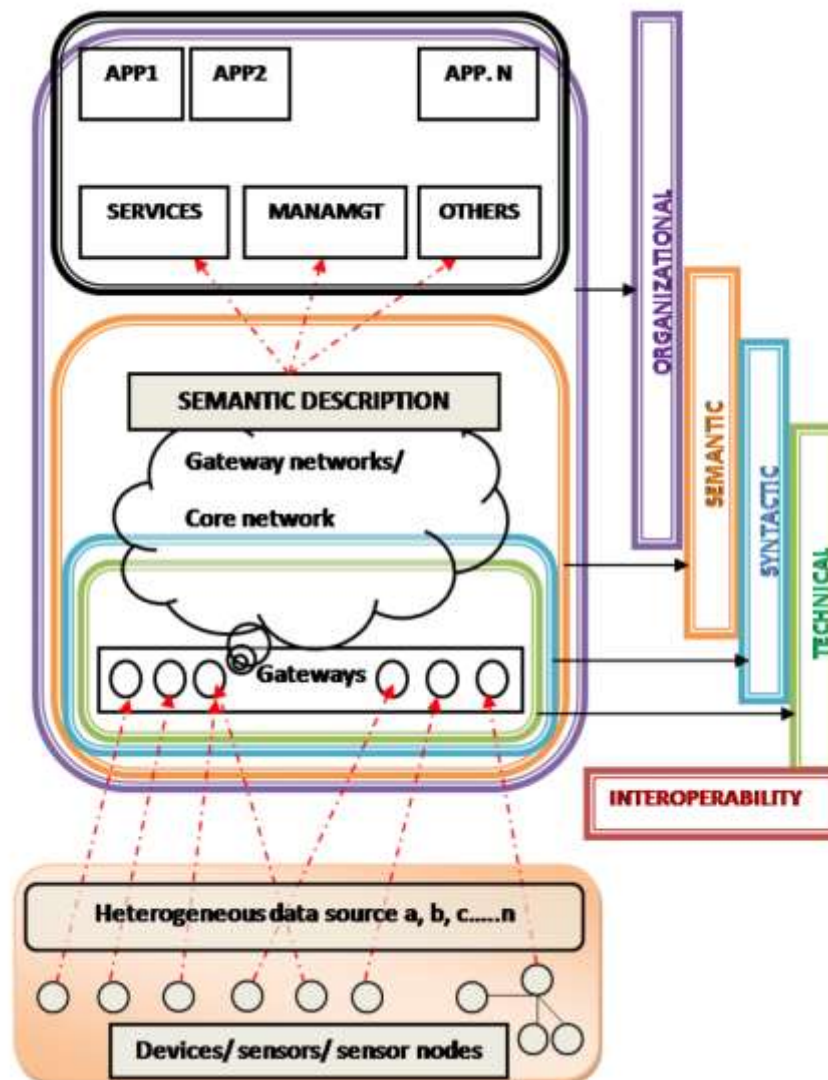


Interoperability solution (1)

- Technical Interoperability: hardware/software level
- Syntactical Interoperability: data format level
- Semantic Interoperability: knowledge level
- Organizational Interoperability: system level



Interoperability solution (2)



Thank you!



上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

上海交通大學

