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

Google 谷歌

阿里云 aliyun.com

DELL EMC

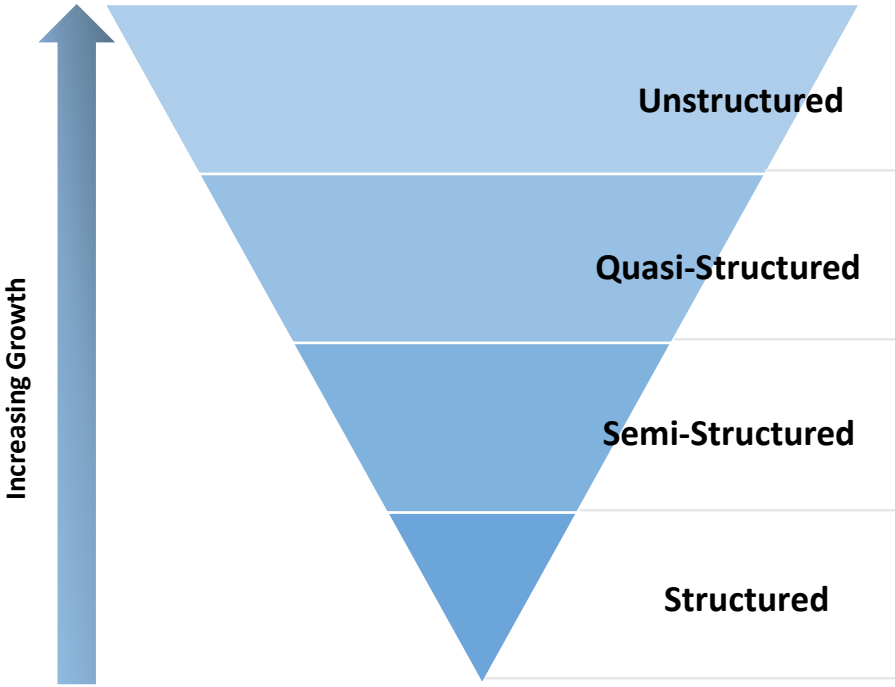
Big Data Definition

- 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



- Data that has no inherent structure and is usually stored as different types of files.
- E.g. Text documents, PDFs, images, and videos

- Textual data with erratic formats that can be formatted with effort and software tools
- E.g. Clickstream data

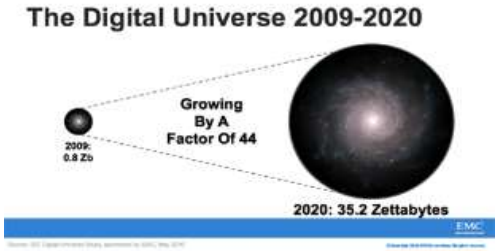
- Textual data files with an apparent pattern, enabling analysis
- E.g. Spreadsheets and XML files

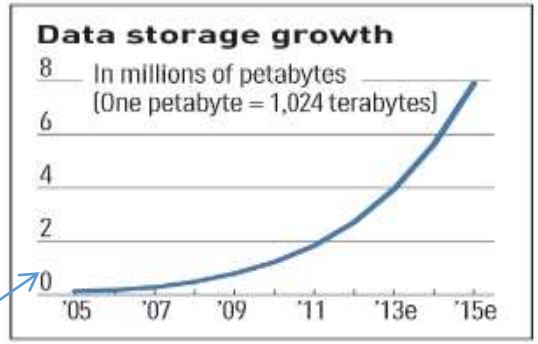
- Data having a defined data model, format, structure
- E.g. Database

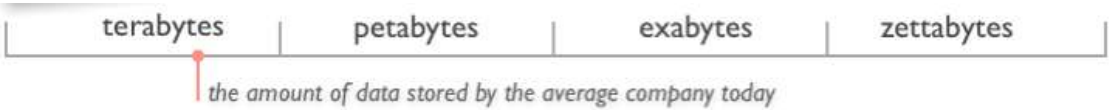
Characteristics of big data

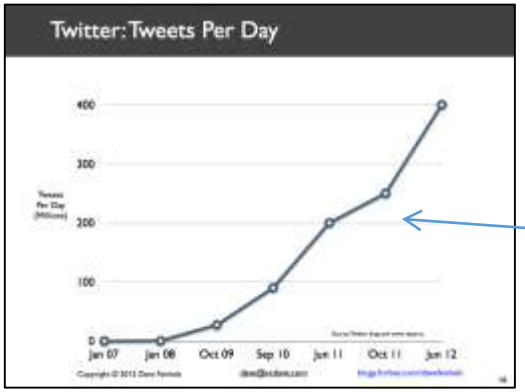
(1-Scale: Volume)

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





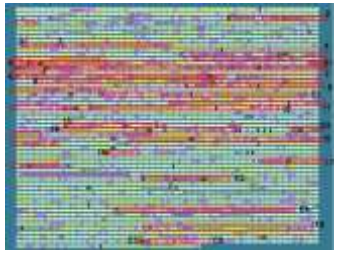


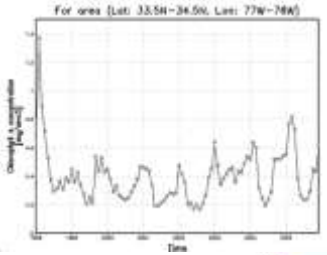


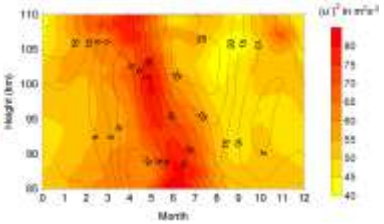
Exponential increase in collected/generated data

Characteristics of big data
(2-Complexity: Varity)

- 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







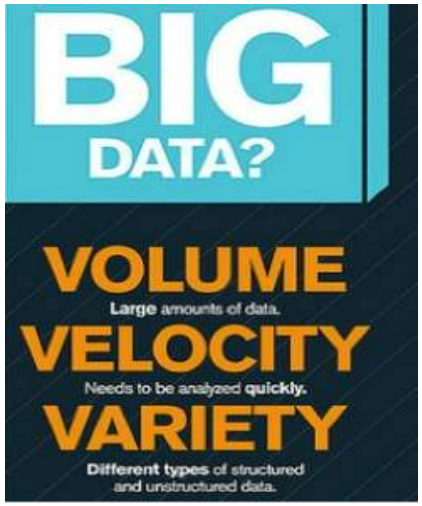
To extract knowledge➔ all these types of data need to linked together

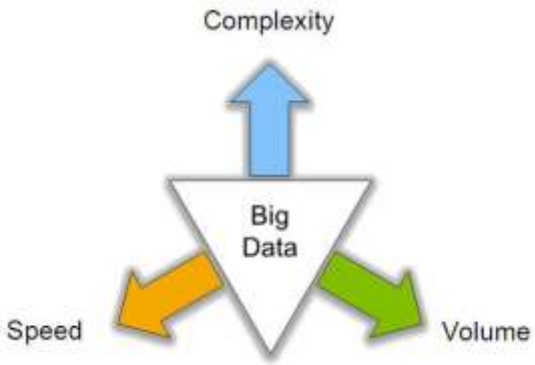
Characteristics of big data
(3-Speed: Velocity)

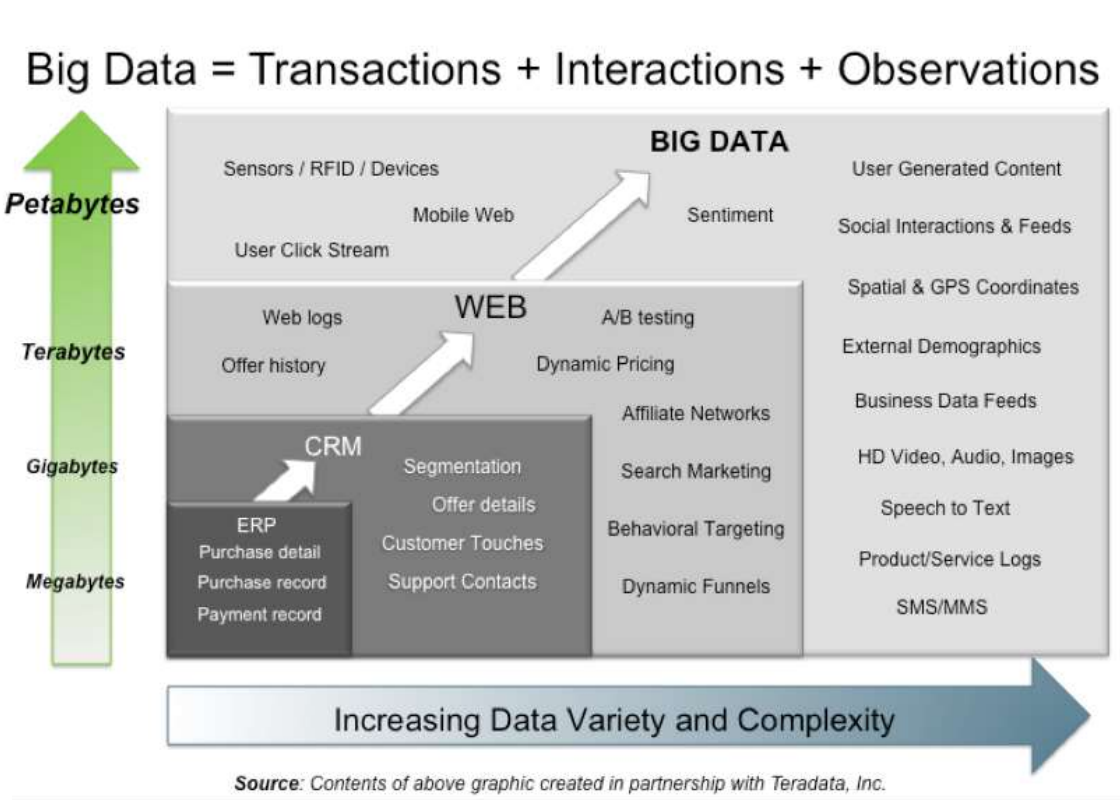
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **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

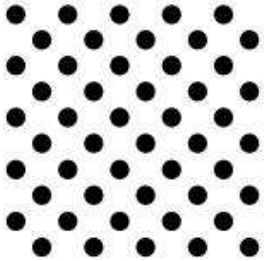
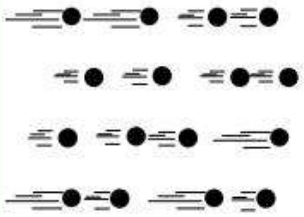
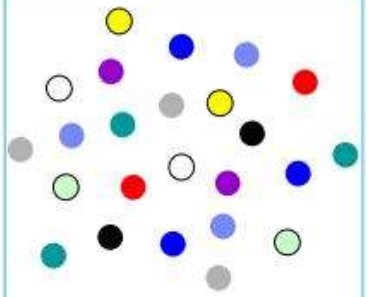
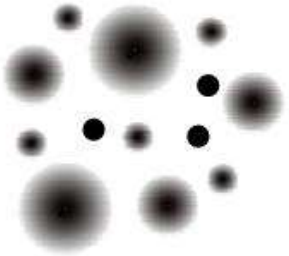
Big Data (3Vs)


BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.


Complexity
Big Data
Speed
Volume


Big Data = Transactions + Interactions + Observations
Petabytes
Terabytes
Gigabytes
Megabytes
BIG DATA
Sensors / RFID / Devices
Mobile Web
User Click Stream
Sentiment
User Generated Content
Social Interactions & Feeds
Spatial & GPS Coordinates
External Demographics
Business Data Feeds
HD Video, Audio, Images
Speech to Text
Product/Service Logs
SMS/MMS
WEB
Web logs
Offer history
A/B testing
Dynamic Pricing
Affiliate Networks
Search Marketing
Behavioral Targeting
Dynamic Funnels
CRM
Segmentation
Offer details
Customer Touches
Support Contacts
ERP
Purchase detail
Purchase record
Payment record
Increasing Data Variety and Complexity
Source: Contents of above graphic created in partnership with Teradata, Inc.

Big Data (4Vs)

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

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

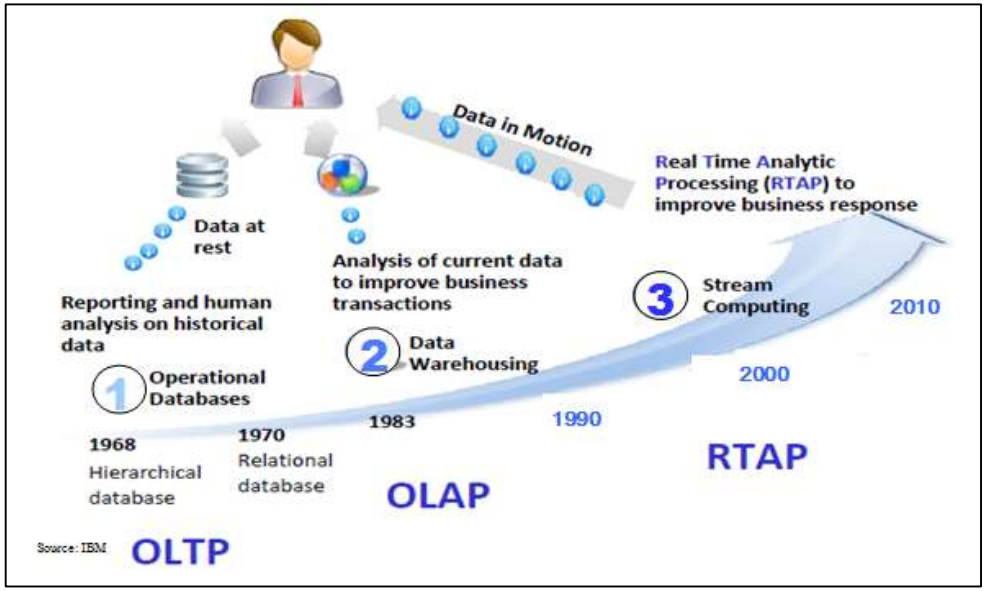
Veracity

- 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

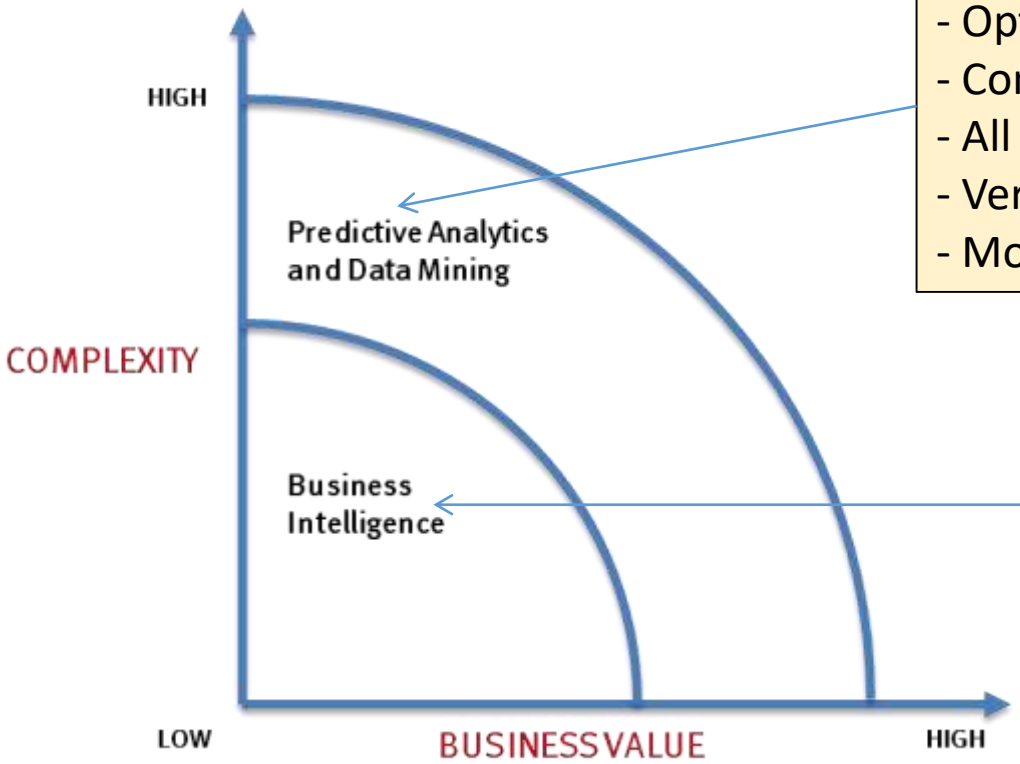
The Model Has Changed…

- **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

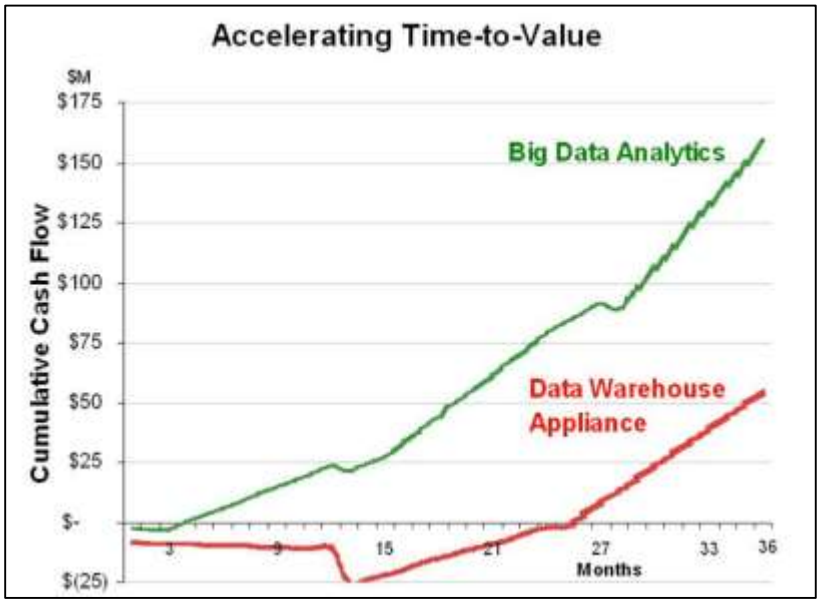


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

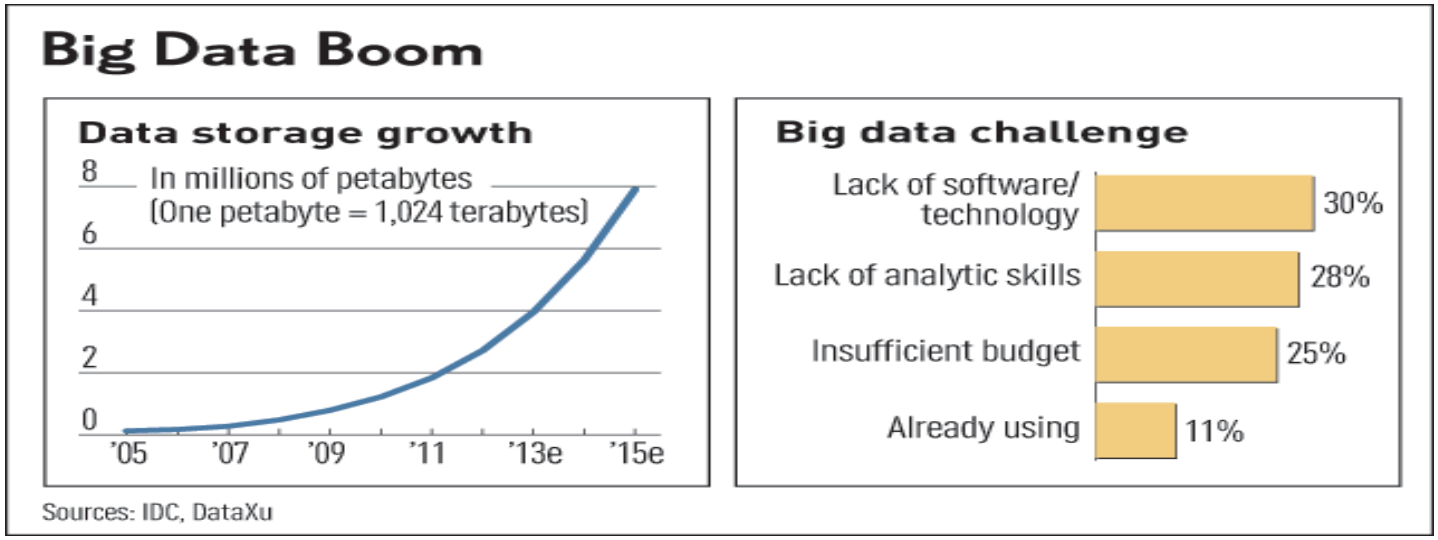
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

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

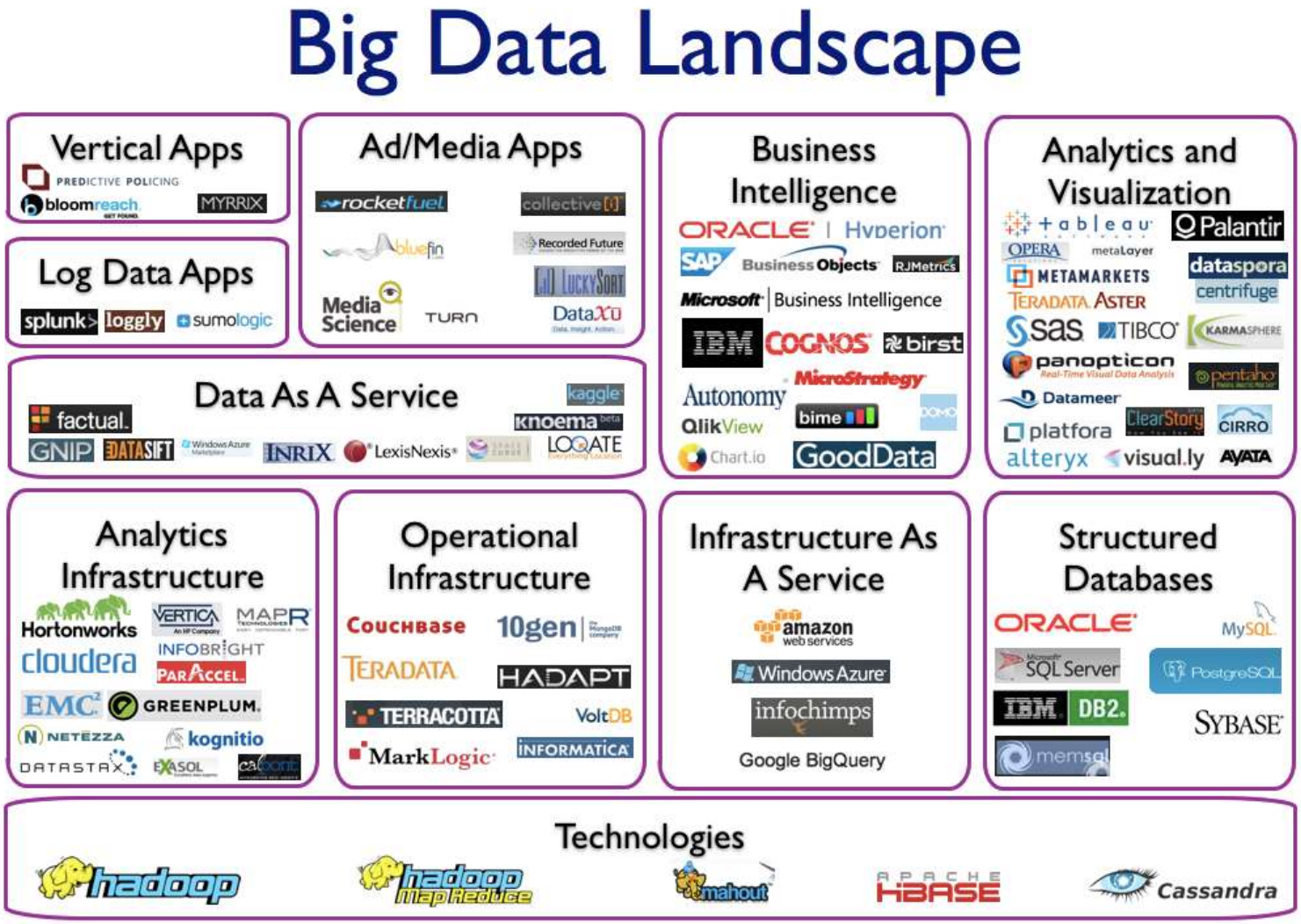


Challenges in Handling Big Data



- **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

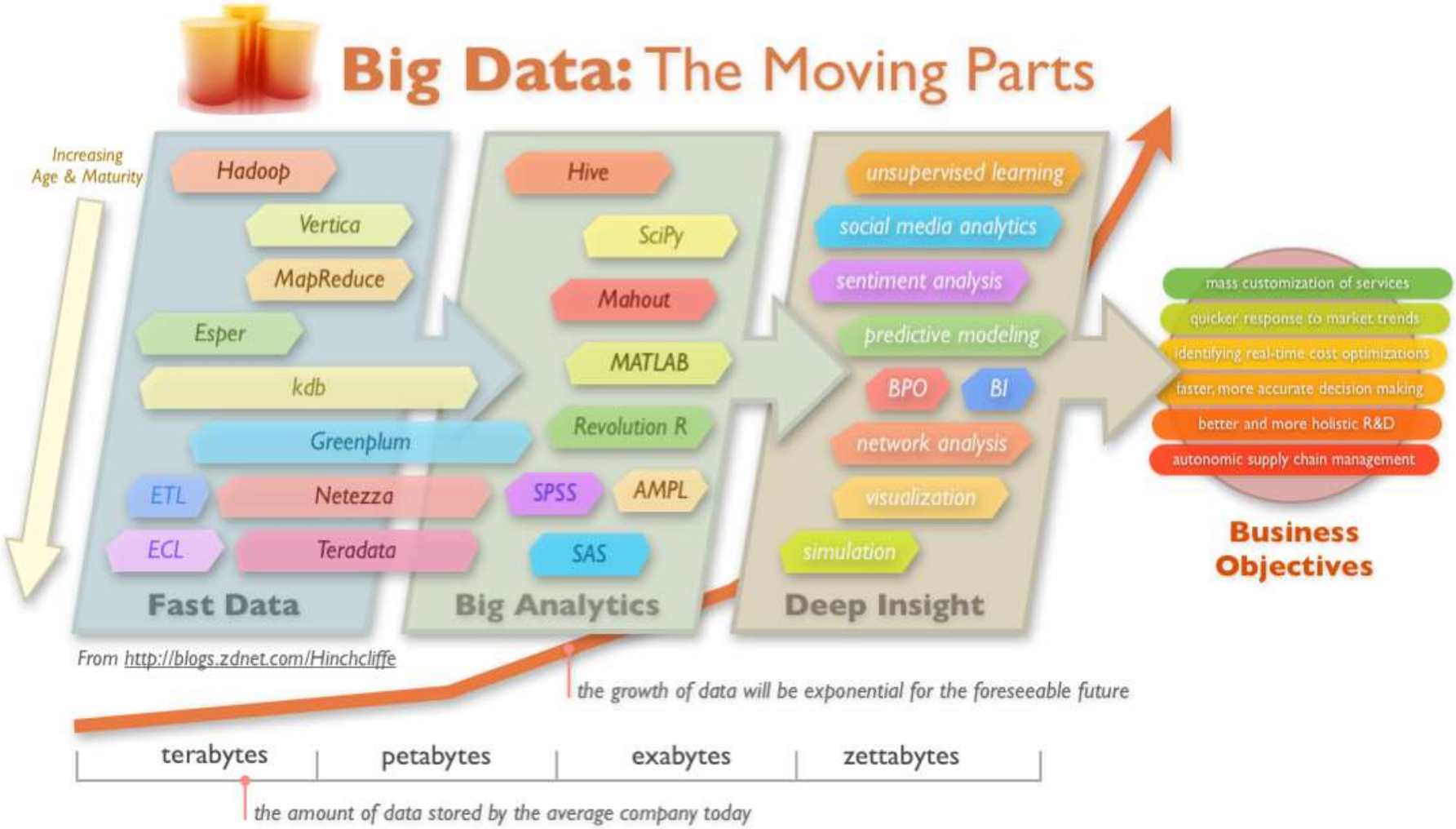
Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Technology



2 Introduction to Cloud Computing

What is Cloud Computing?

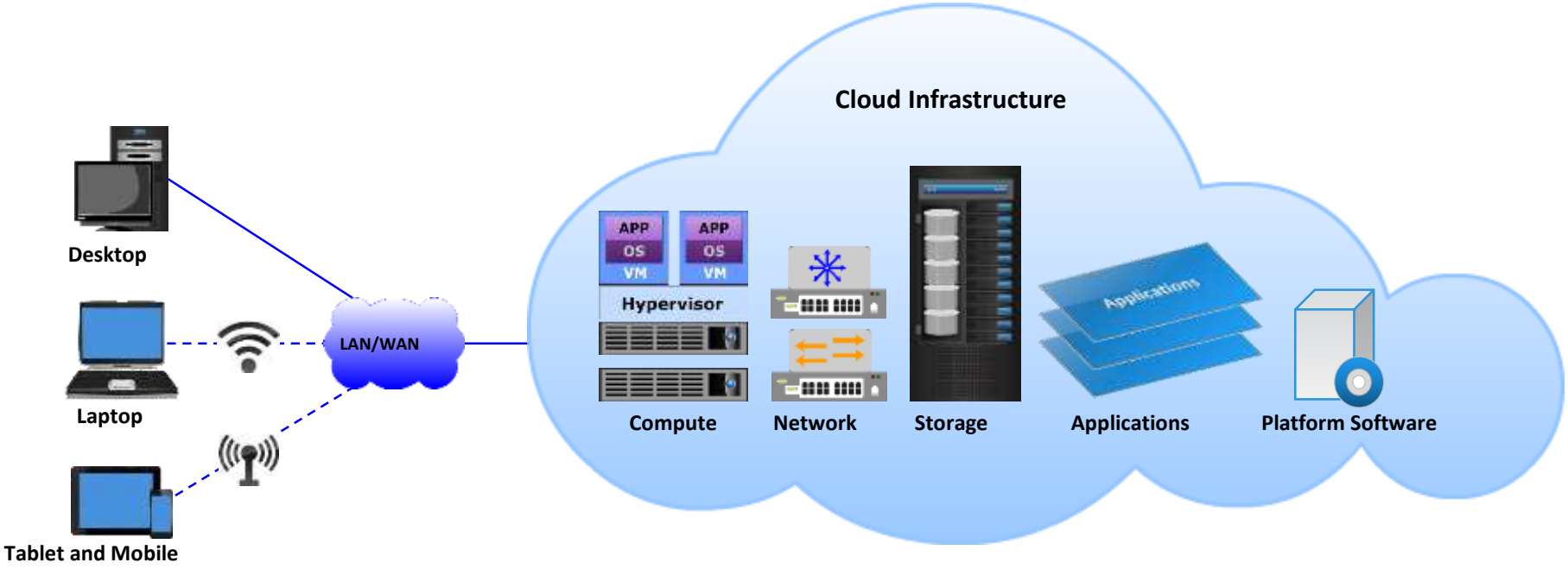
Cloud Computing

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.

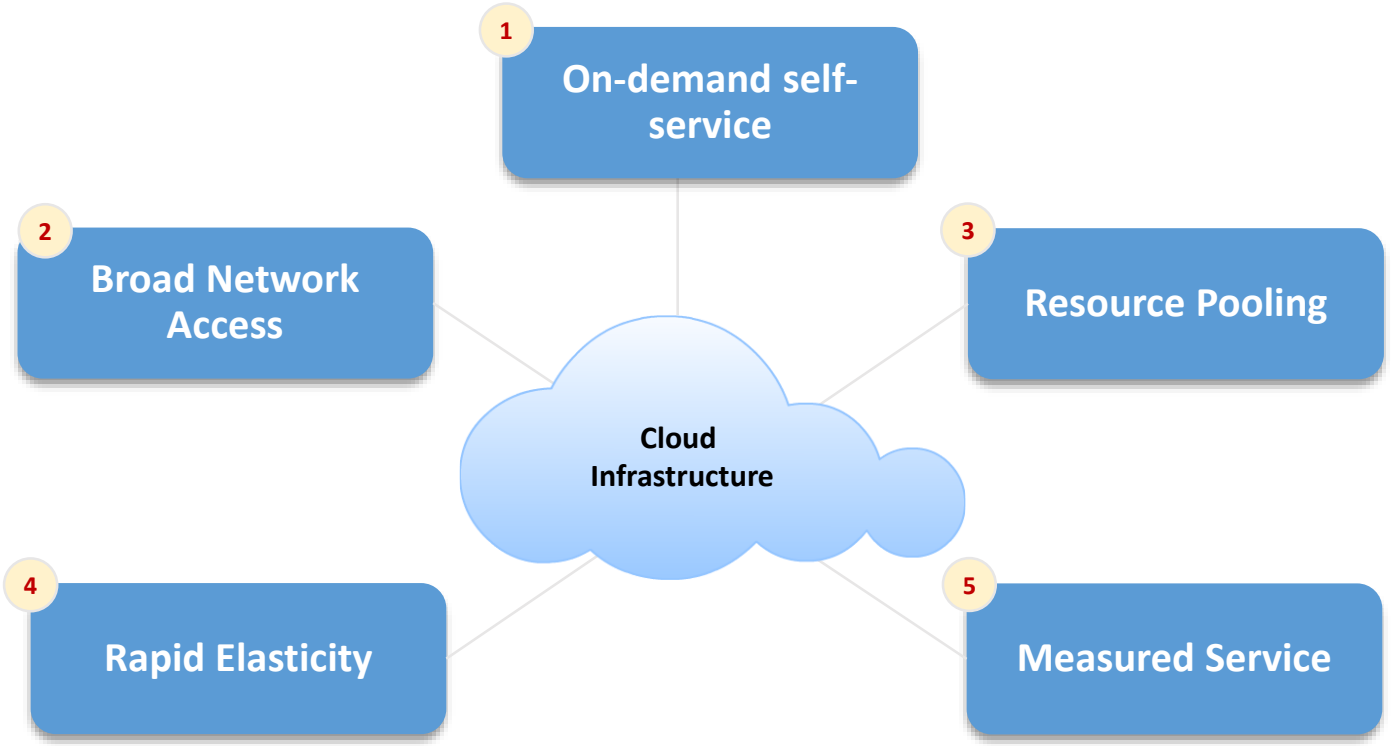
– U.S. National Institute of Standards and Technology, Special Publication 800-145

- 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

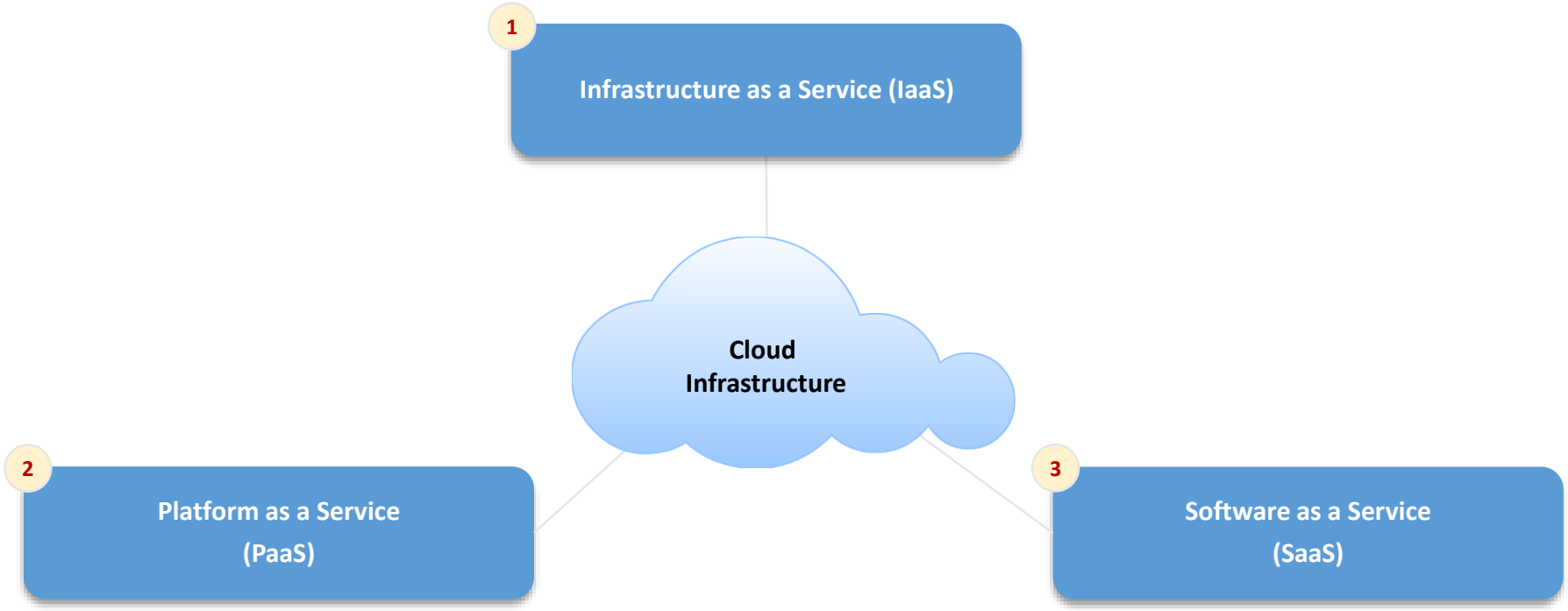
What is Cloud Computing? (Cont'd)



Essential Cloud Characteristics



Cloud Service Models

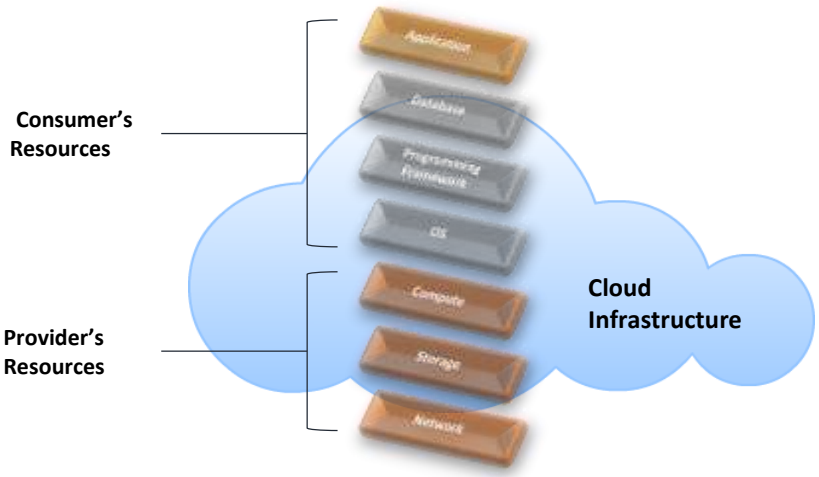


Infrastructure as a Service

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

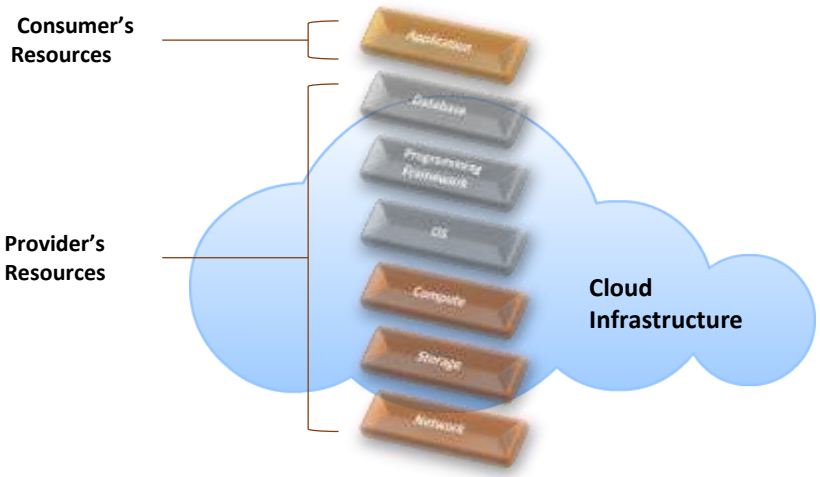


Platform as a Service

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

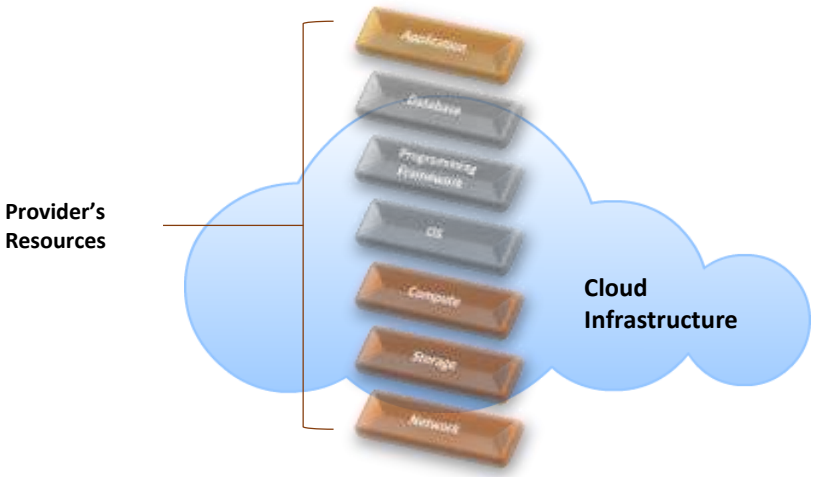


Software as a Service

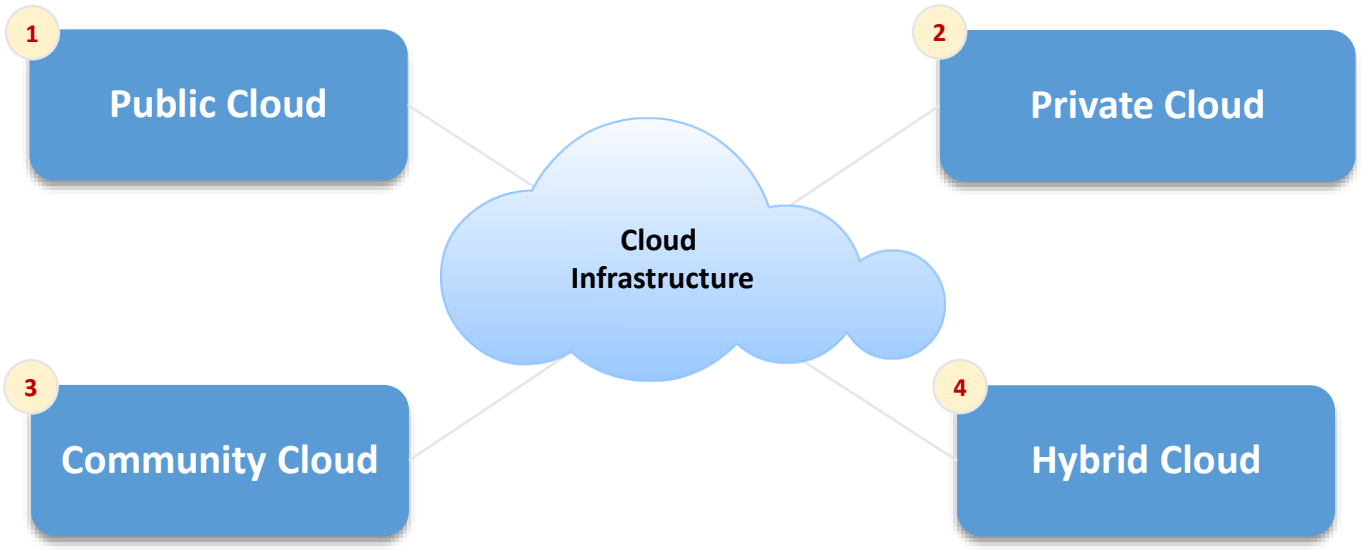
Software as a Service

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.

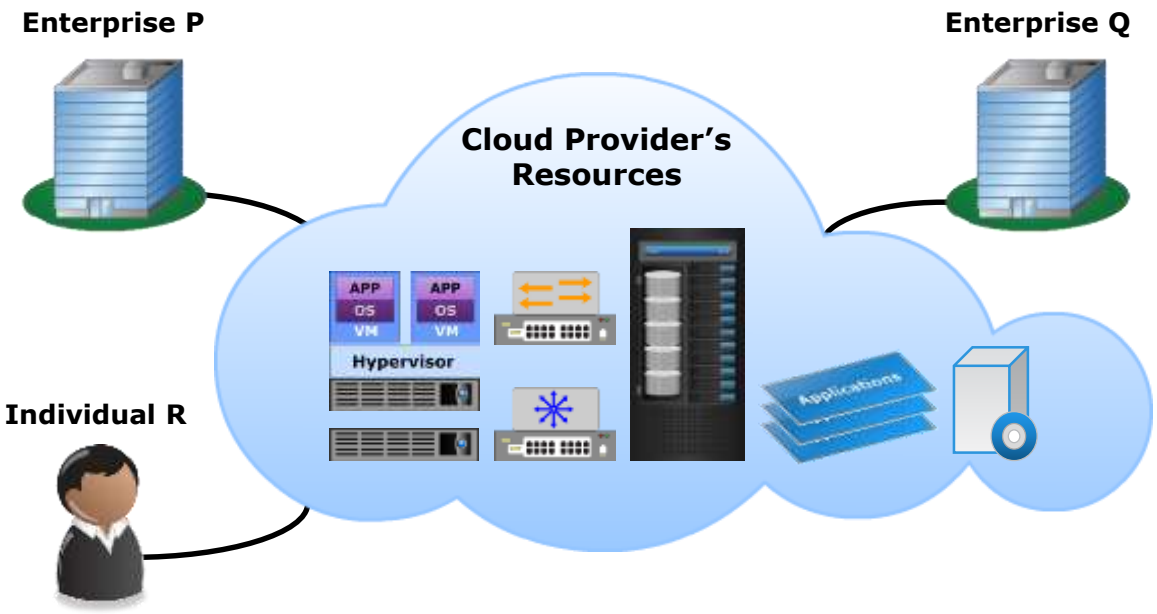
– U.S. National Institute of Standards and Technology, Special Publication 800-145



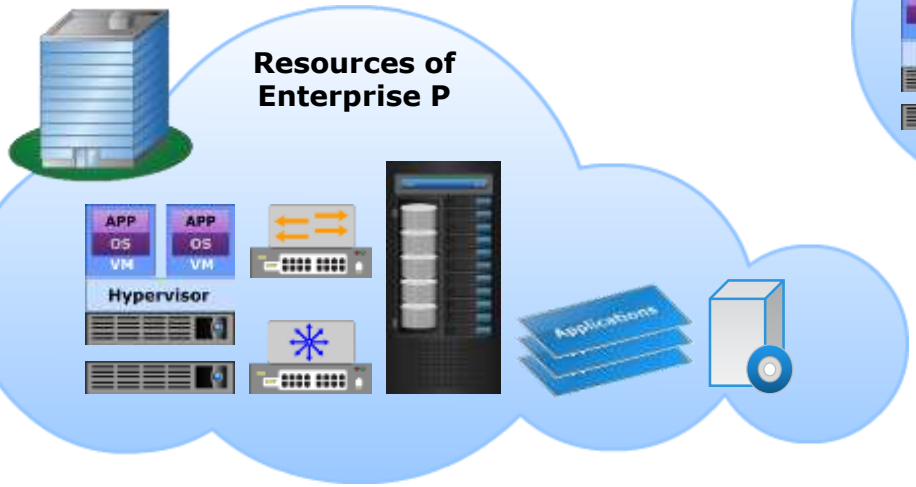
Cloud Deployment Models



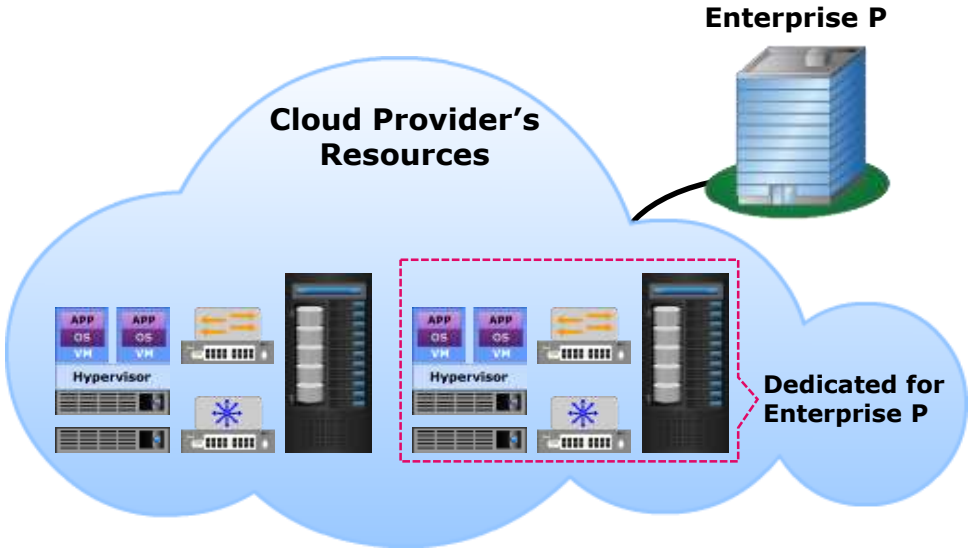
Public Cloud



Private Cloud



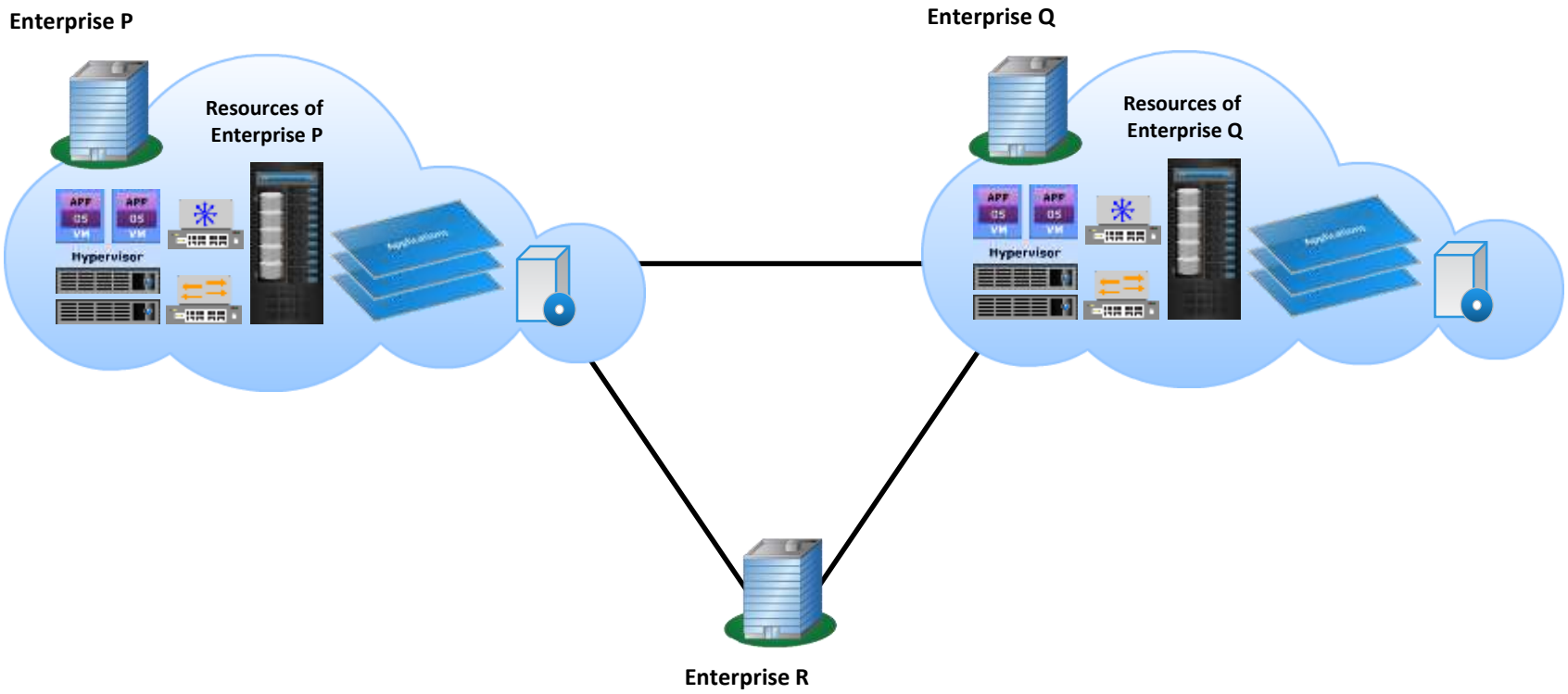
1) On-premise Private Cloud



2) Externally-hosted Private Cloud

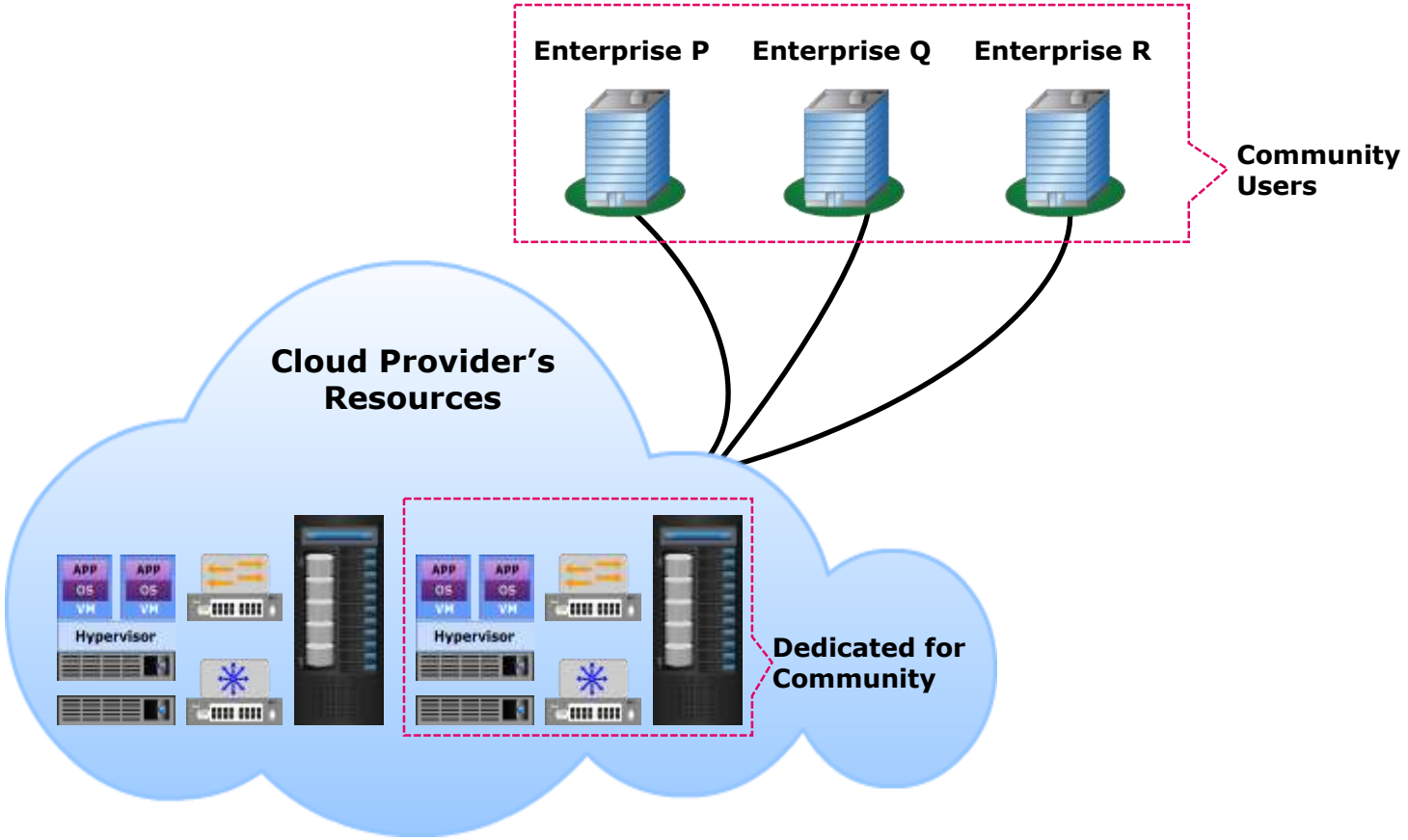
Community Cloud

- On-premise Community Cloud

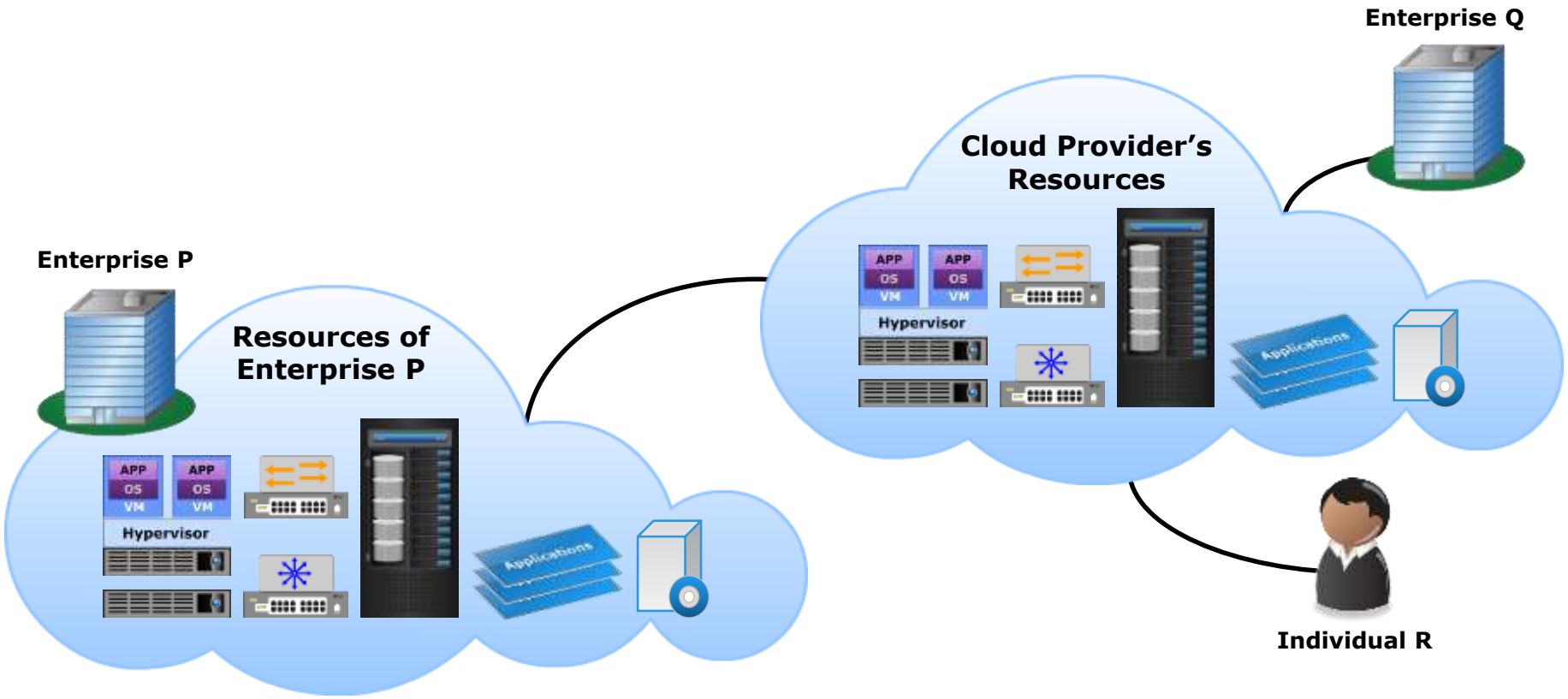


Community Cloud

- Externally-hosted Community Cloud



Hybrid Cloud



3

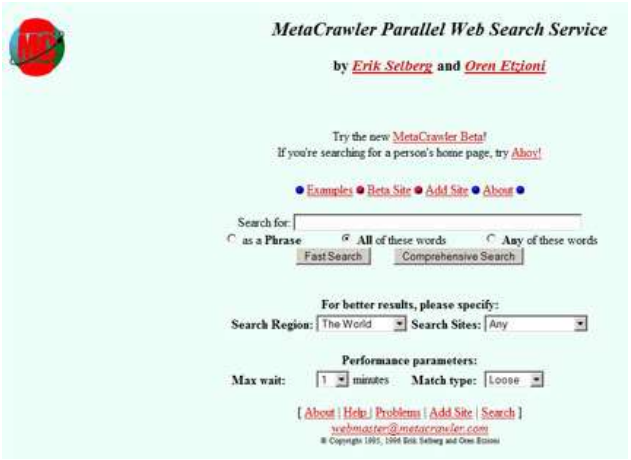
Industrial Solutions

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



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

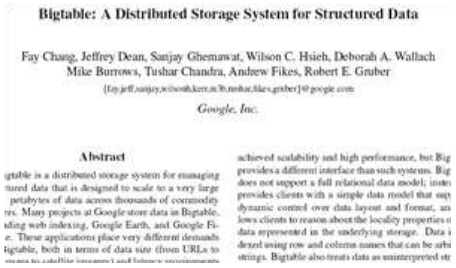
MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

2006



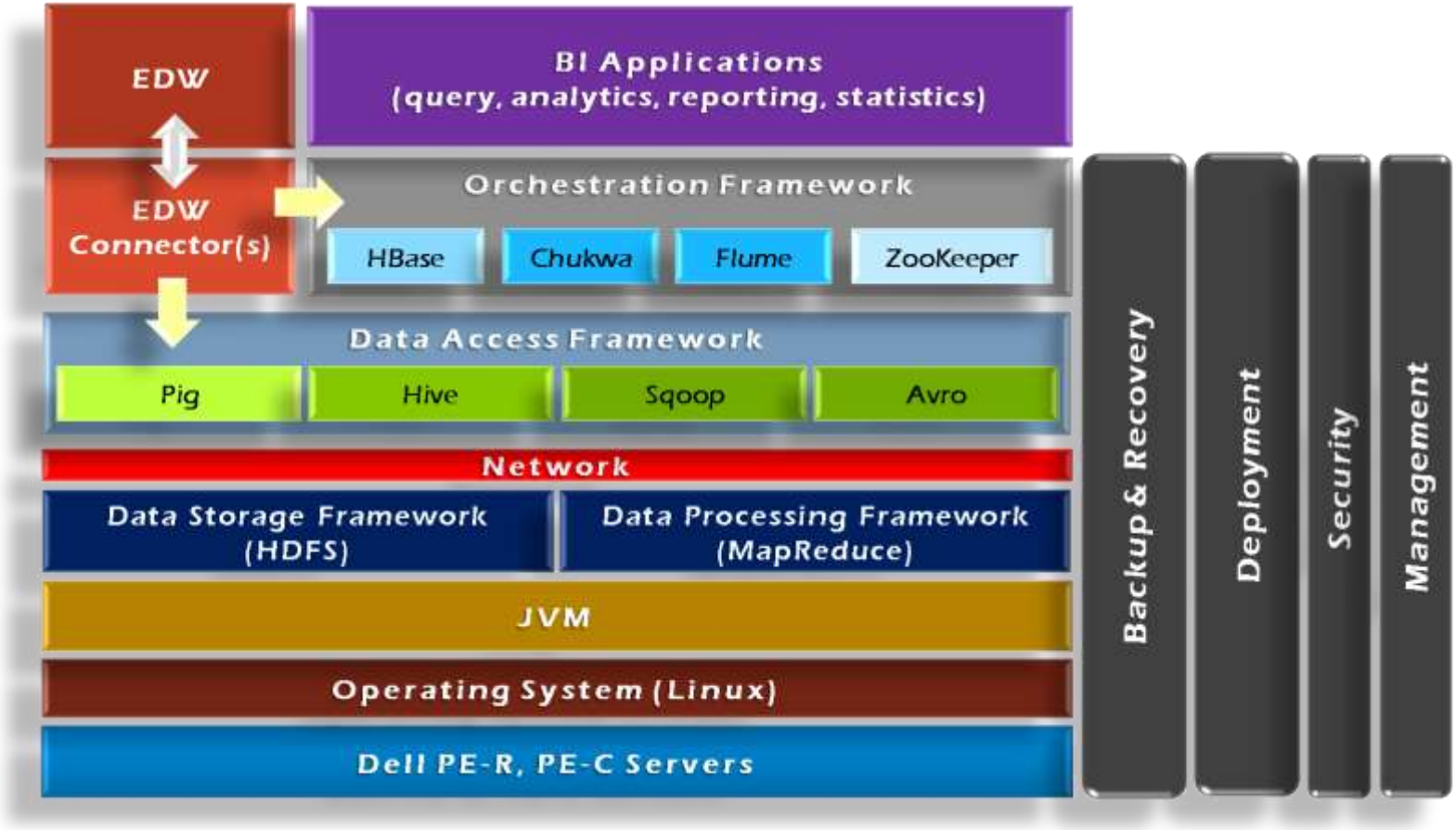
Bigtable: A Distributed Storage System for Structured Data

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

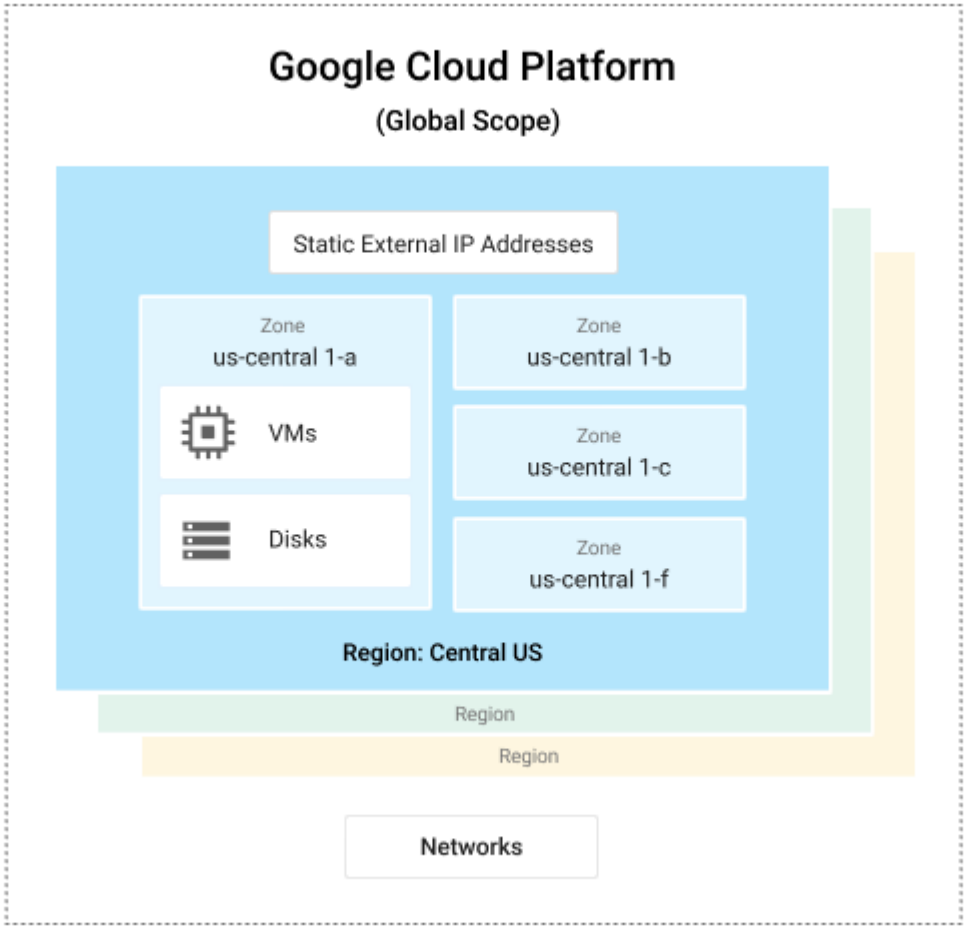
Google, Inc.

Abstract

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

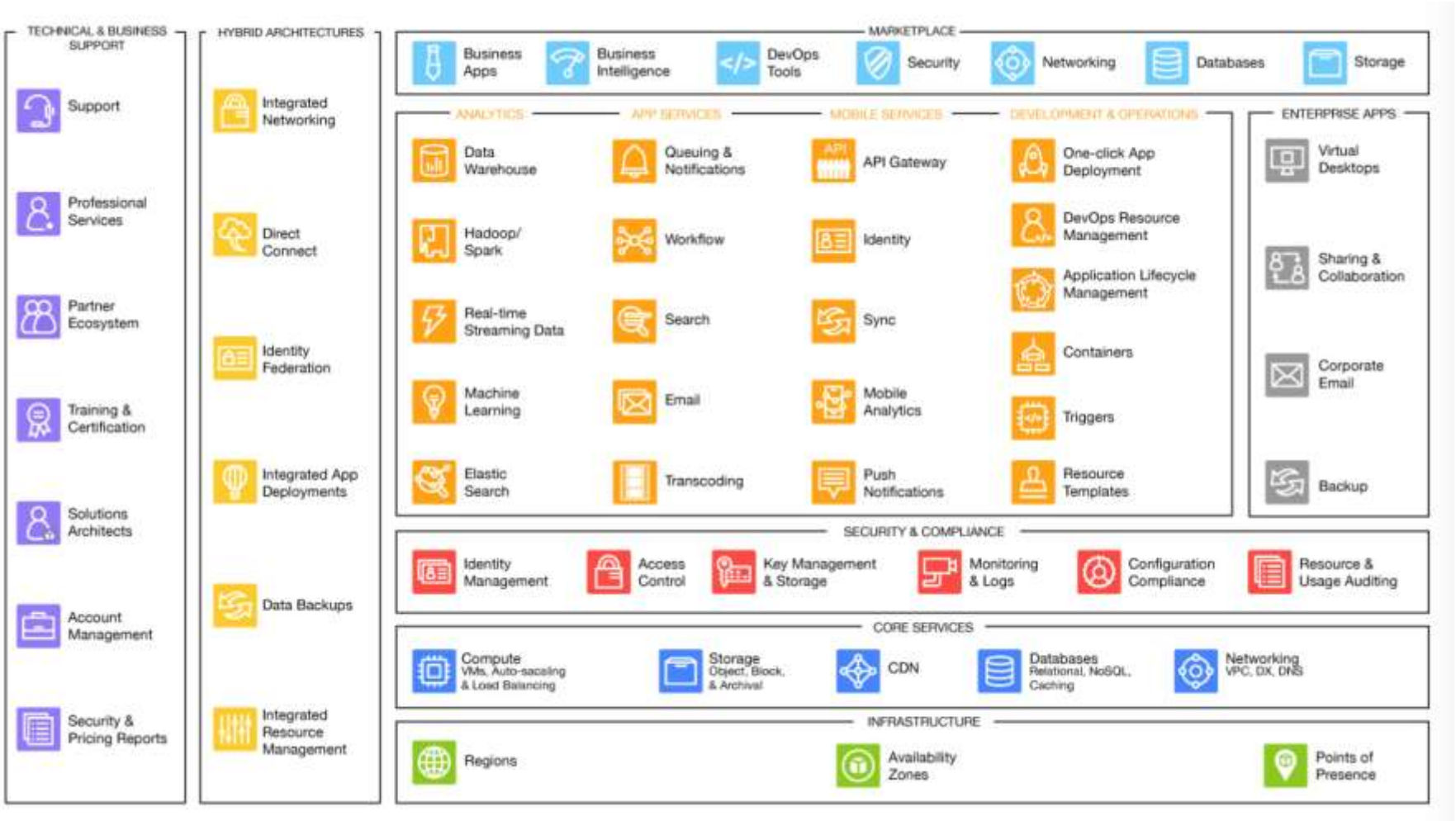
Amazon

- 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 web services™

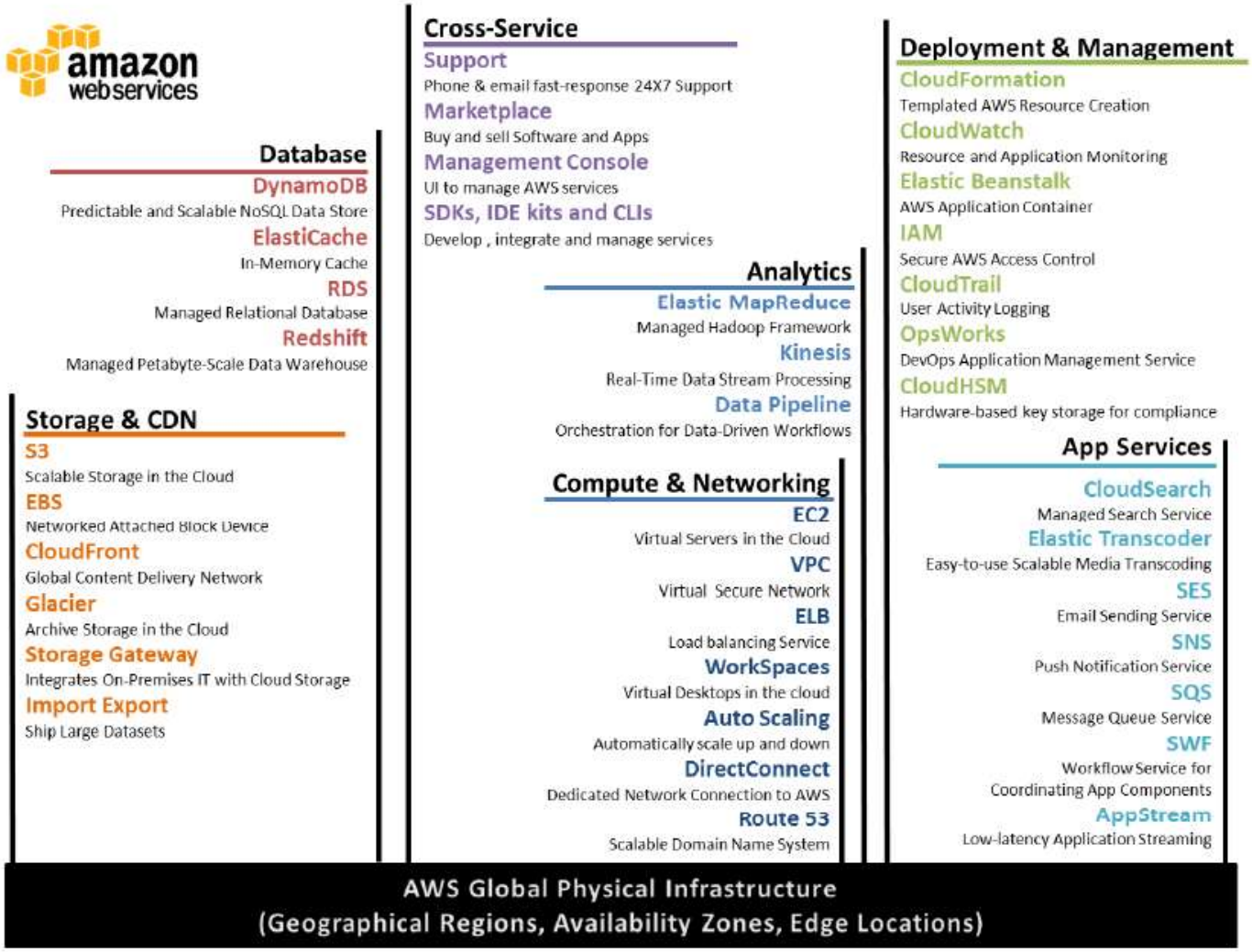
AWS



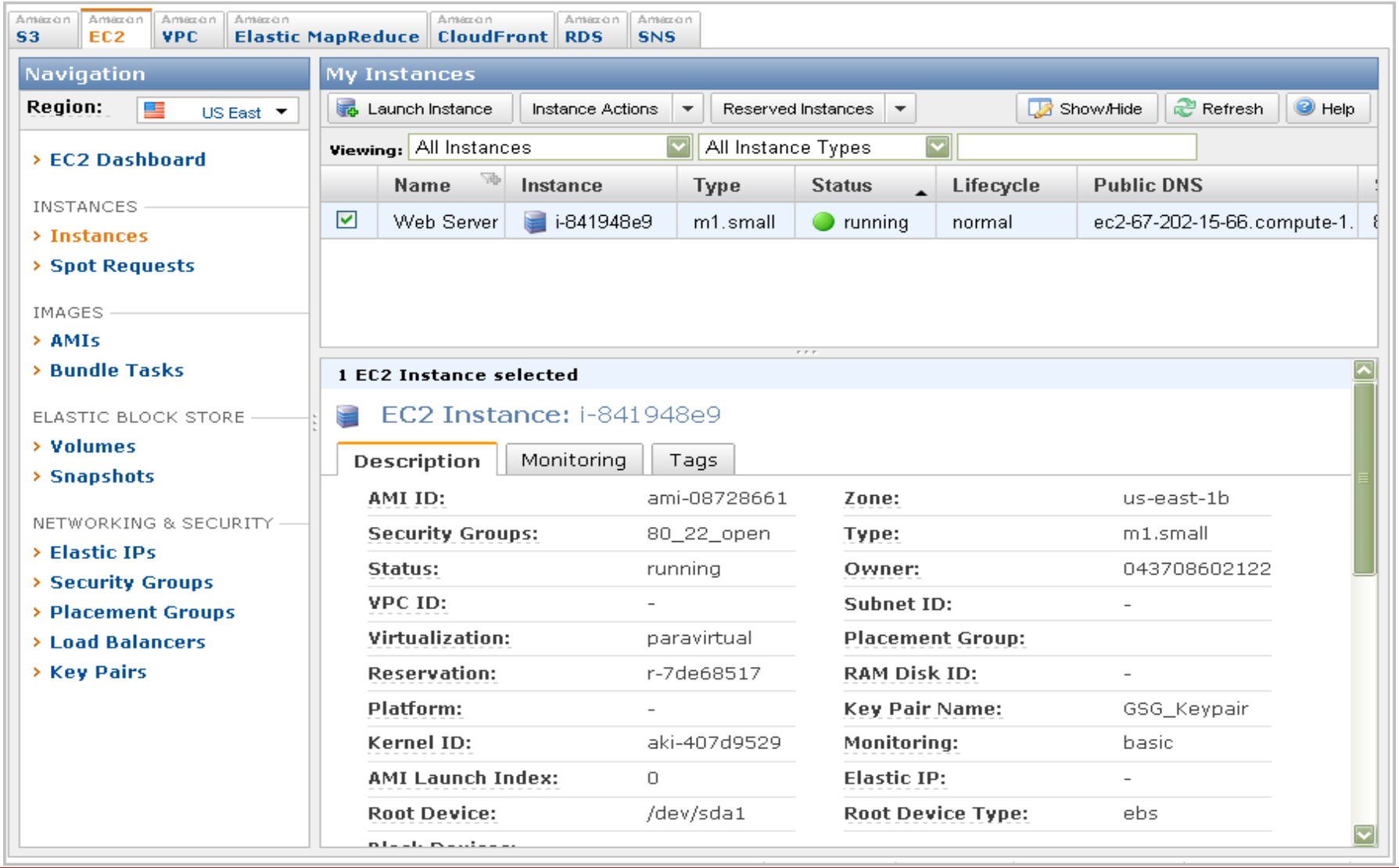
Amazon

amazon web services™

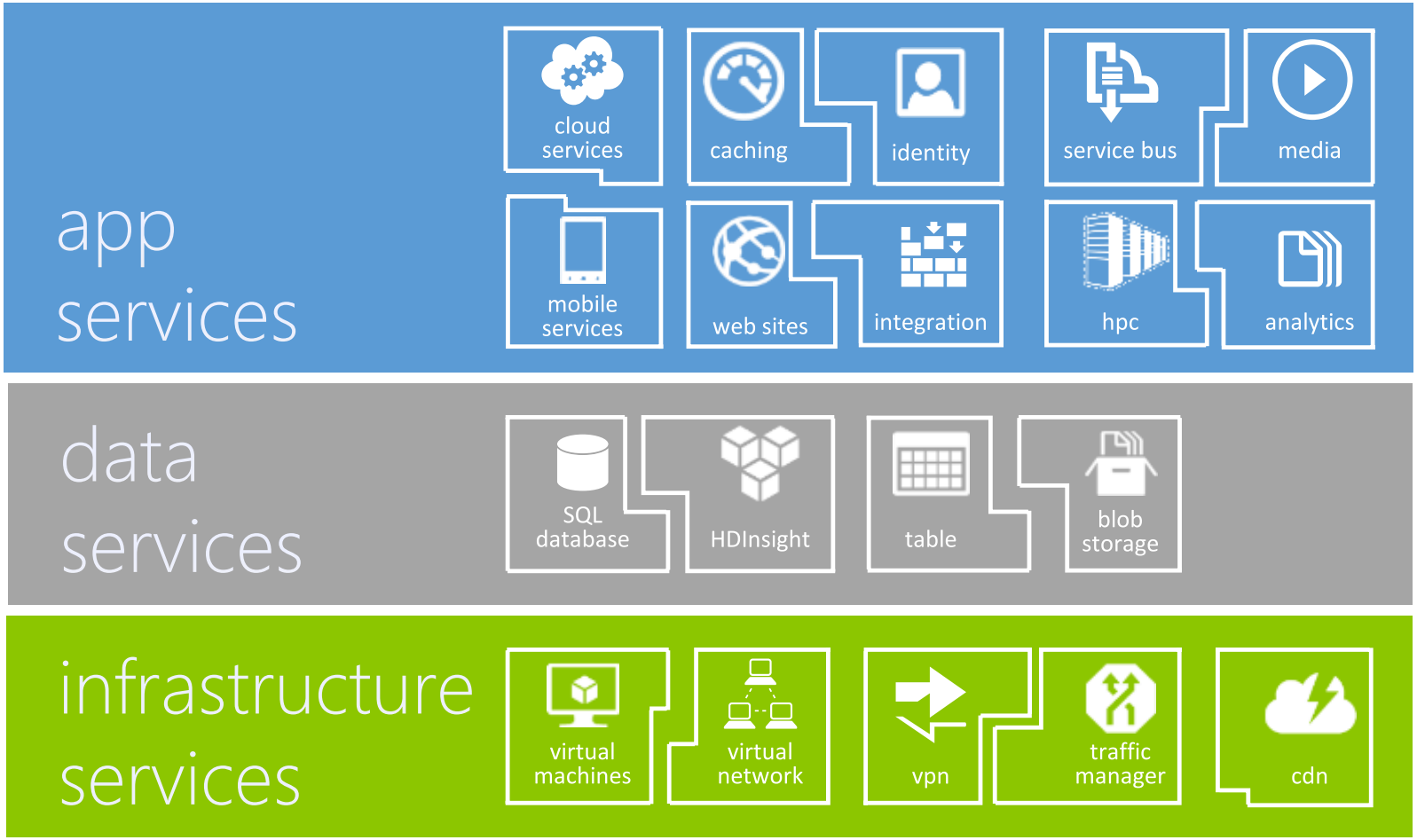
AWS



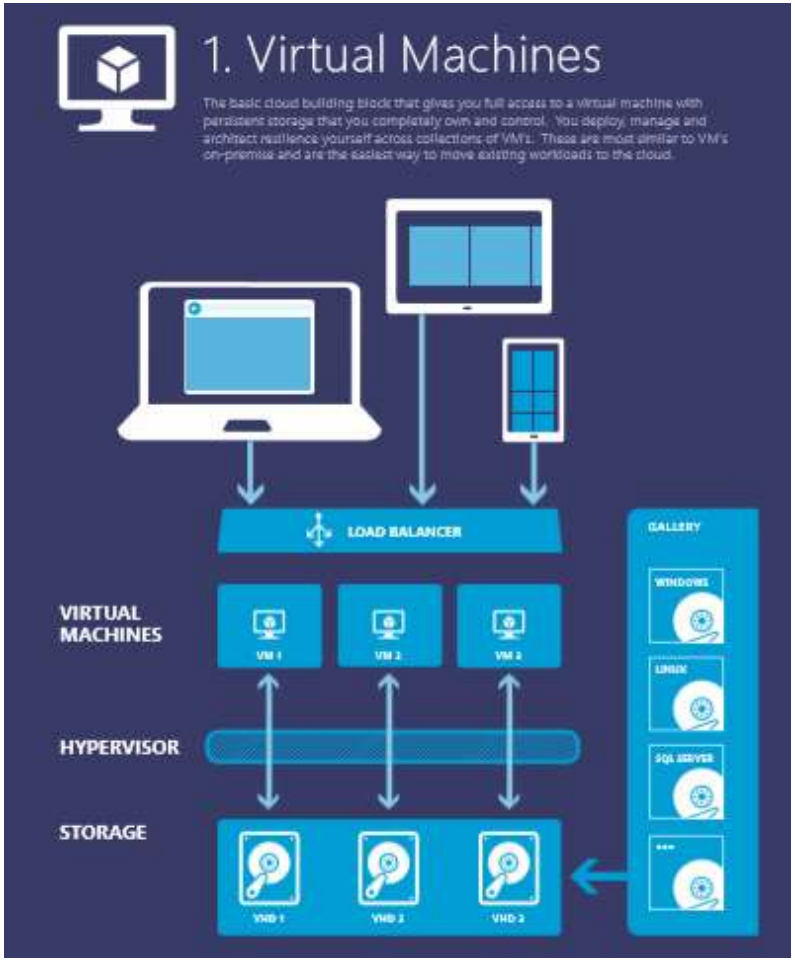
AWS Management Console

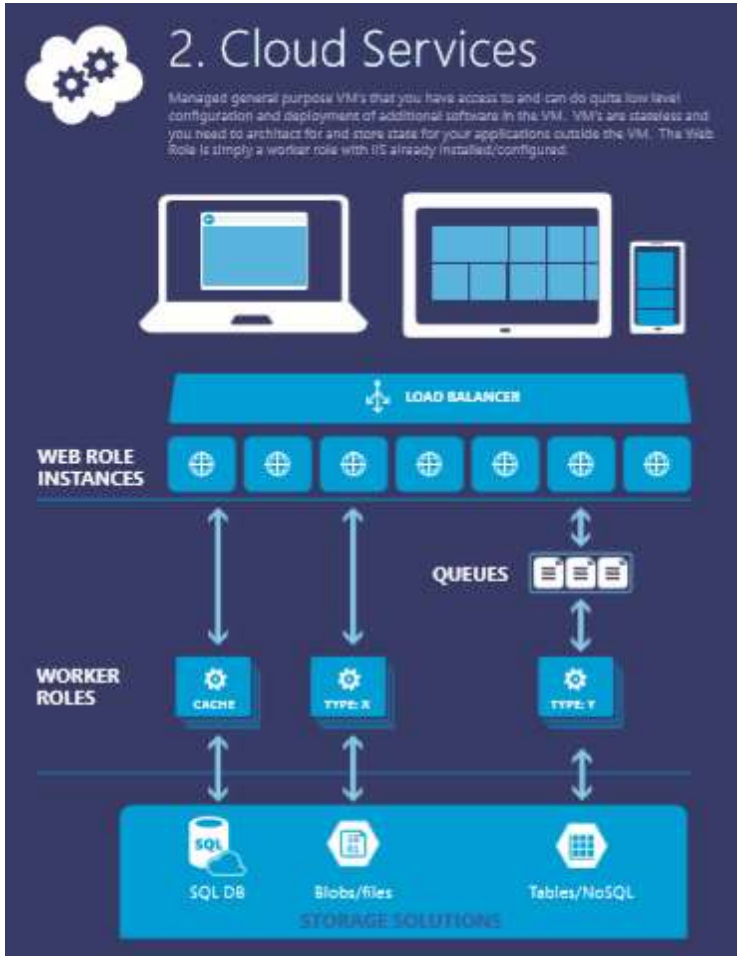


Microsoft Azure (1) Microsoft Azure

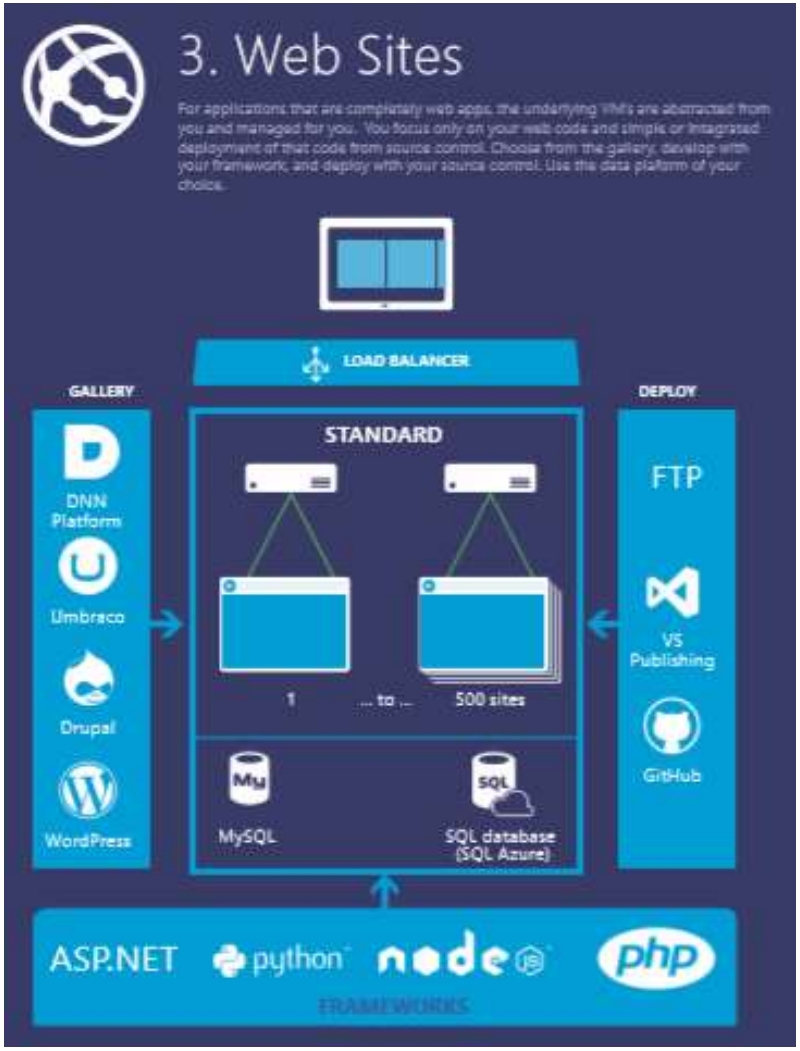


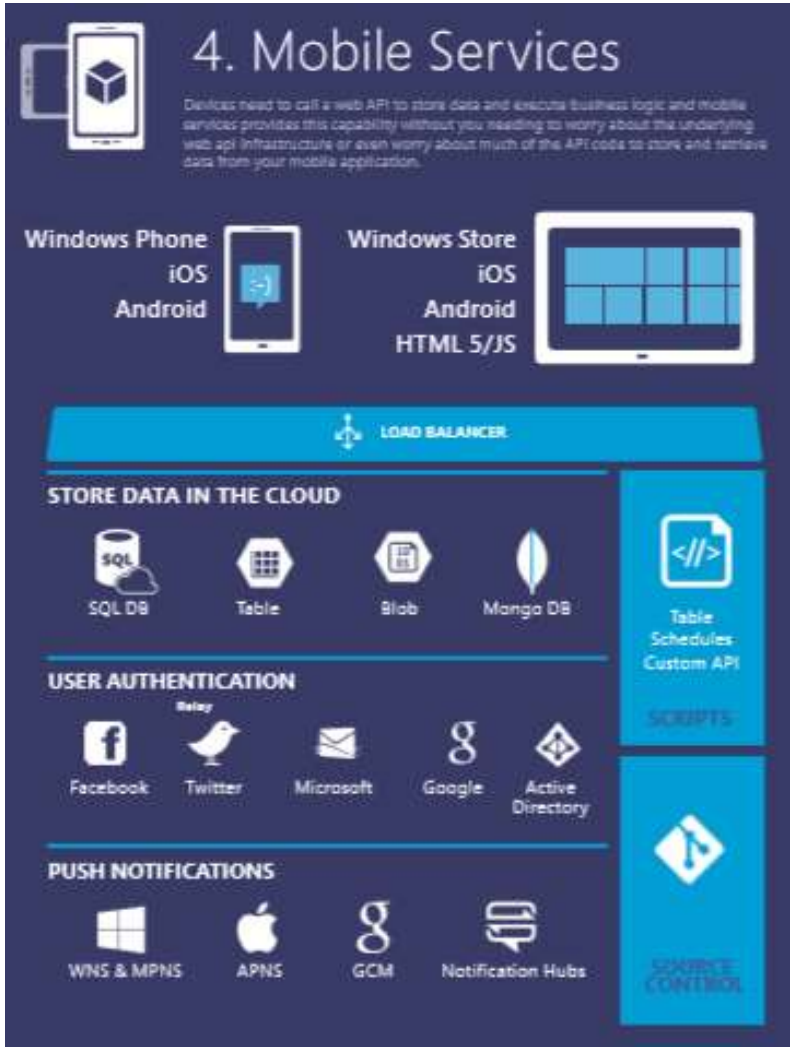
Microsoft Azure (2) Microsoft Azure





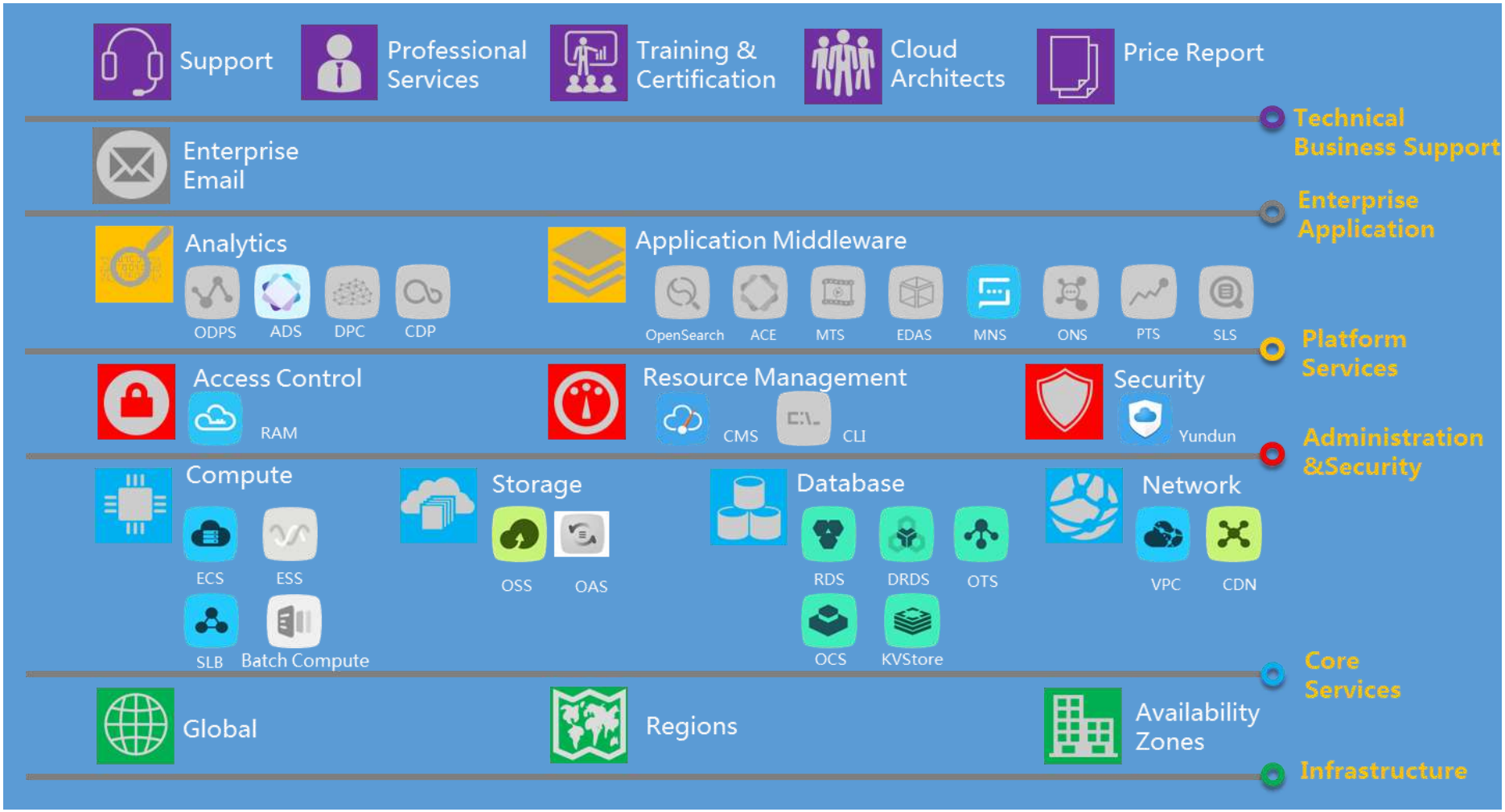
Microsoft Azure (3) Microsoft Azure





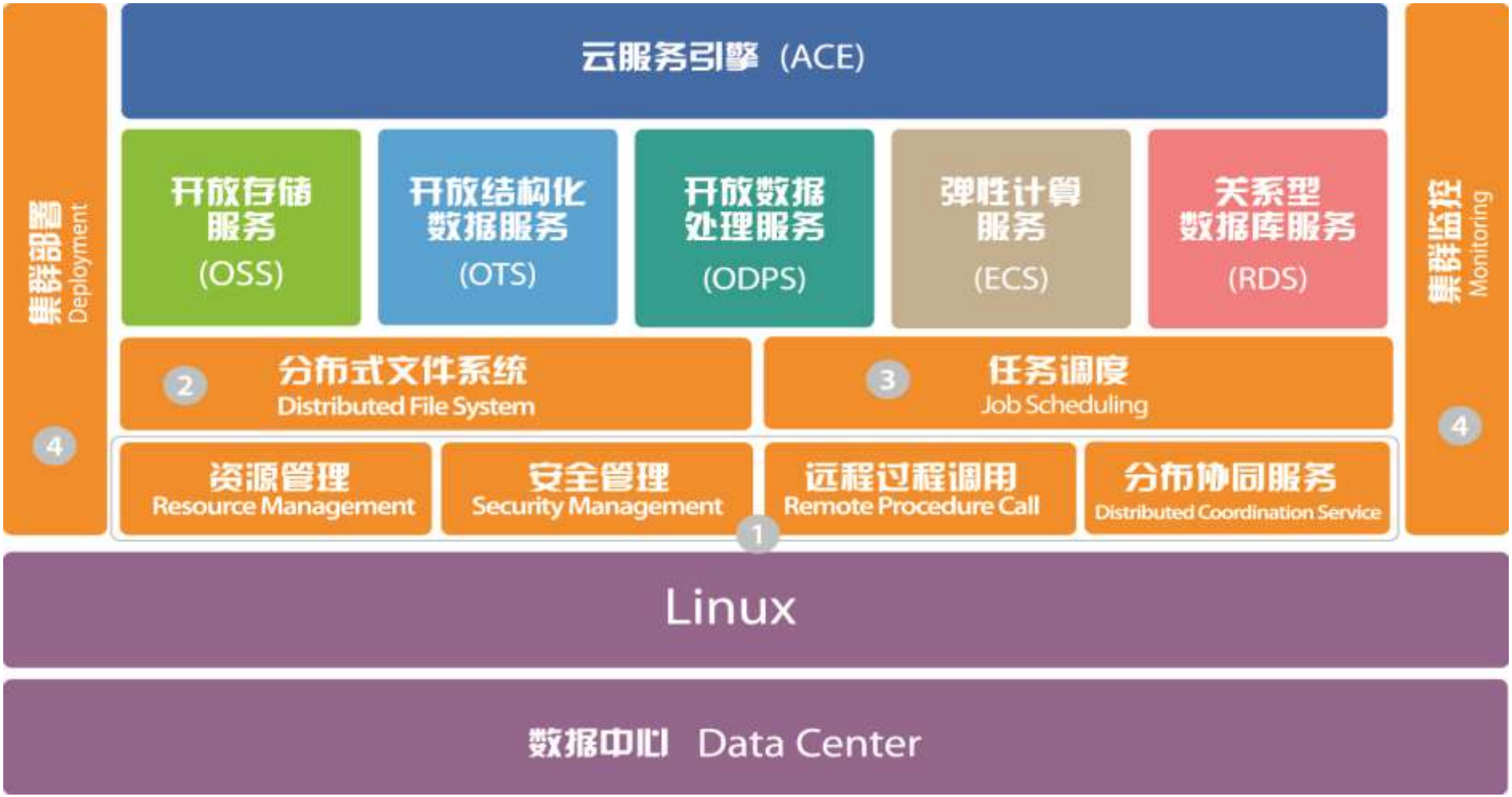
Aliyun Framework(1)

阿里云 aliyun.com



Aliyun Framework (2)

阿里云 aliyun.com



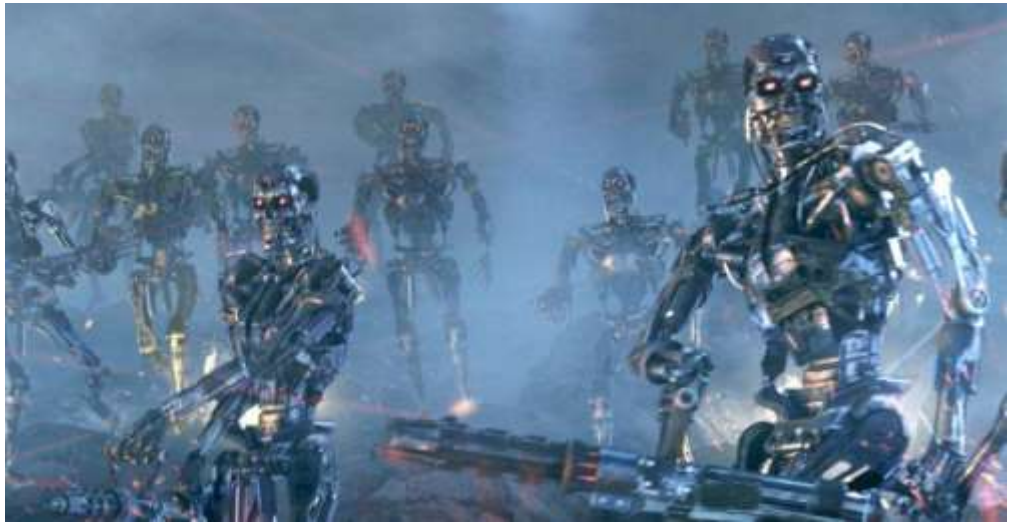
4 IoT

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

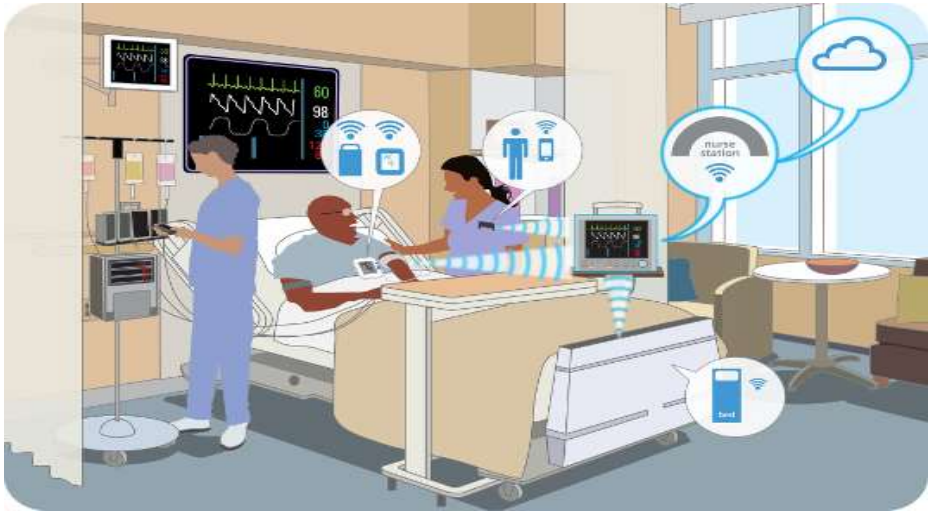
- M2M (Machine to Machine)
- “Internet of Everything” (Cisco Systems)
- “World Size Web” (Bruce Schneier)
- “Skynet” (Terminator movie)



Where is IoT

Wearable Tech

Smart Appliances



Healthcare

IoT Access Many Industries

Healthcare and Life Sciences

Municipal Infrastructure

Smart Home

Retail

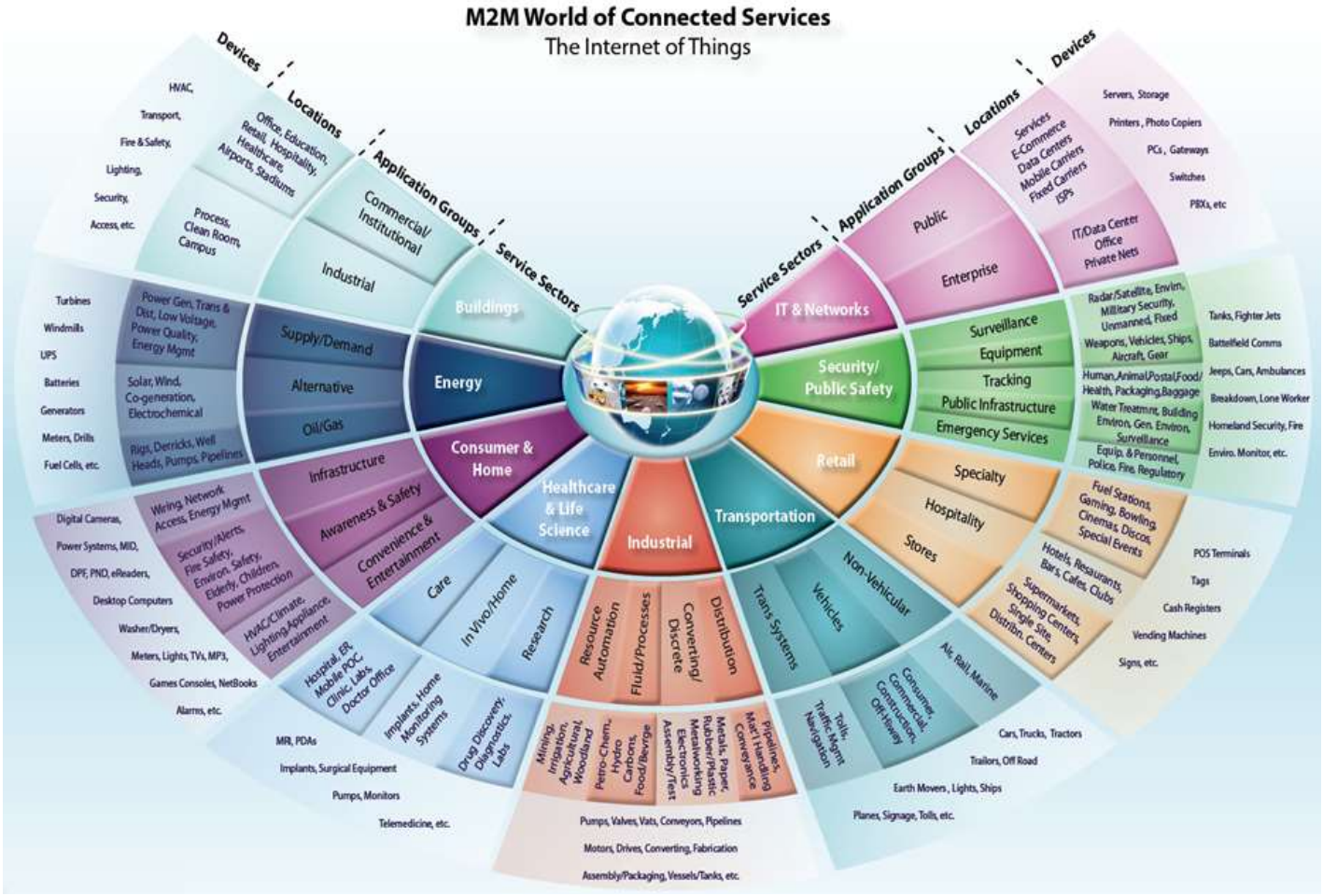
Manufacturing, Logistics & Supply Chain

Agriculture

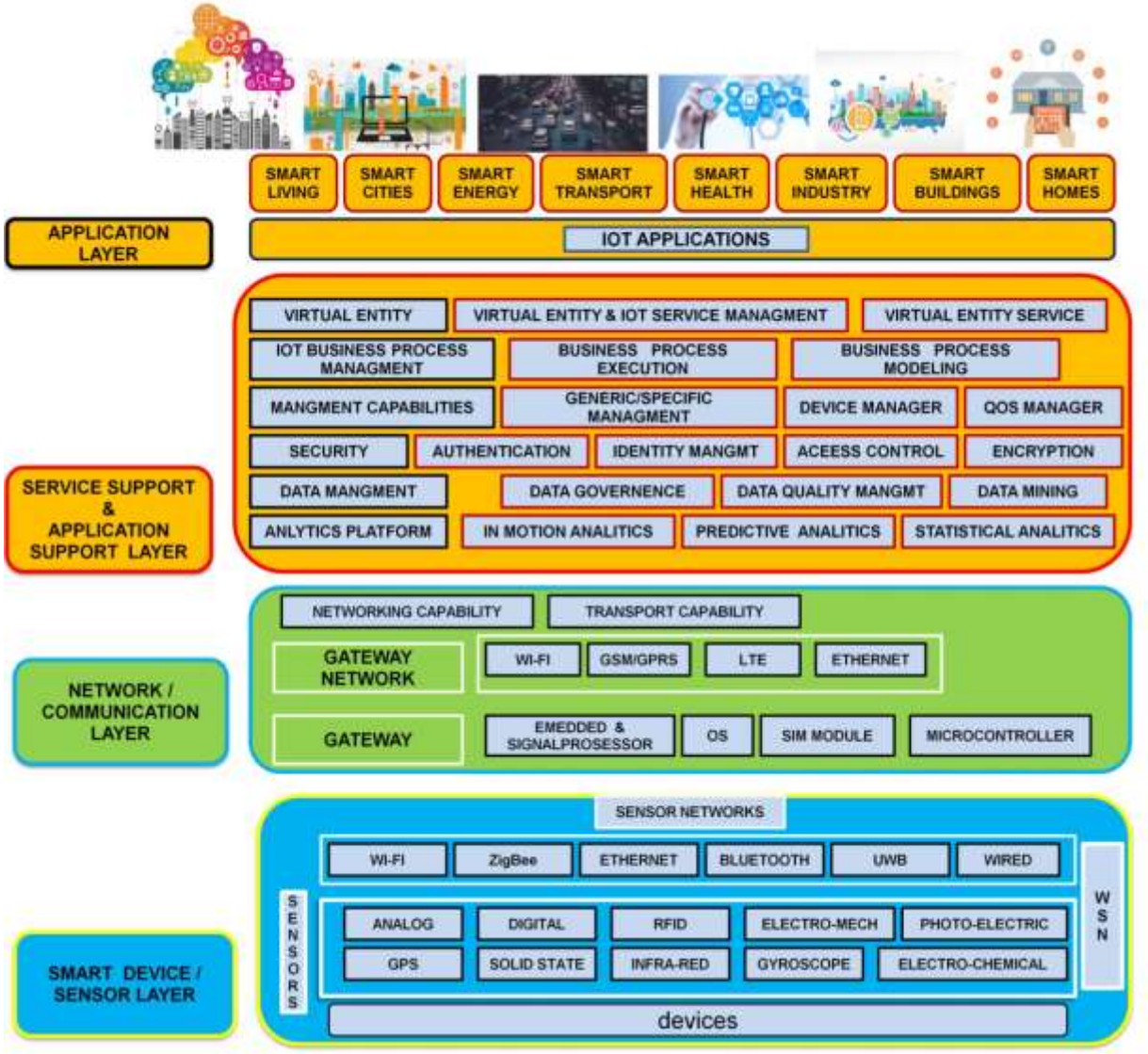
Education

Automotive

IoT Ecosystem

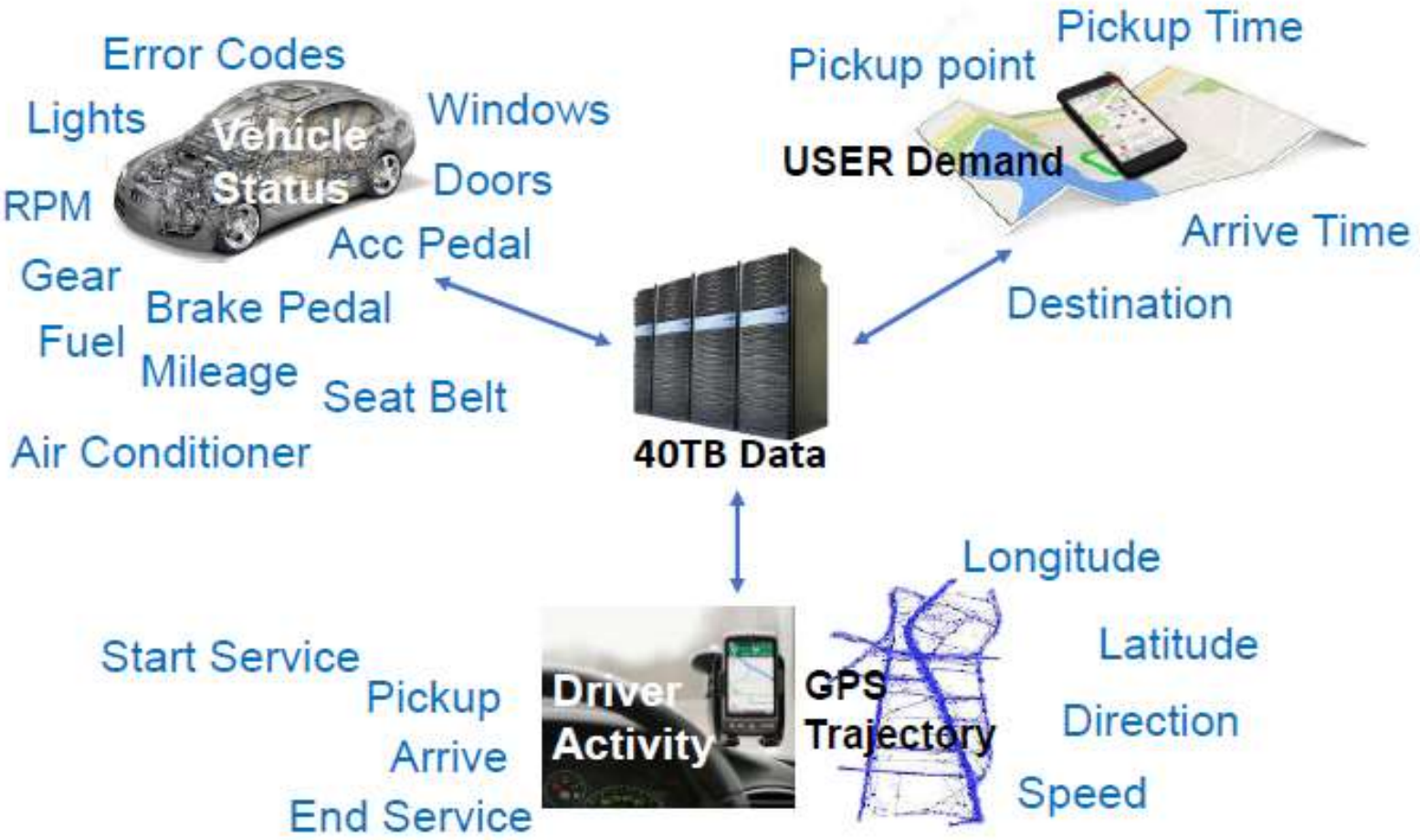


IoT Integration

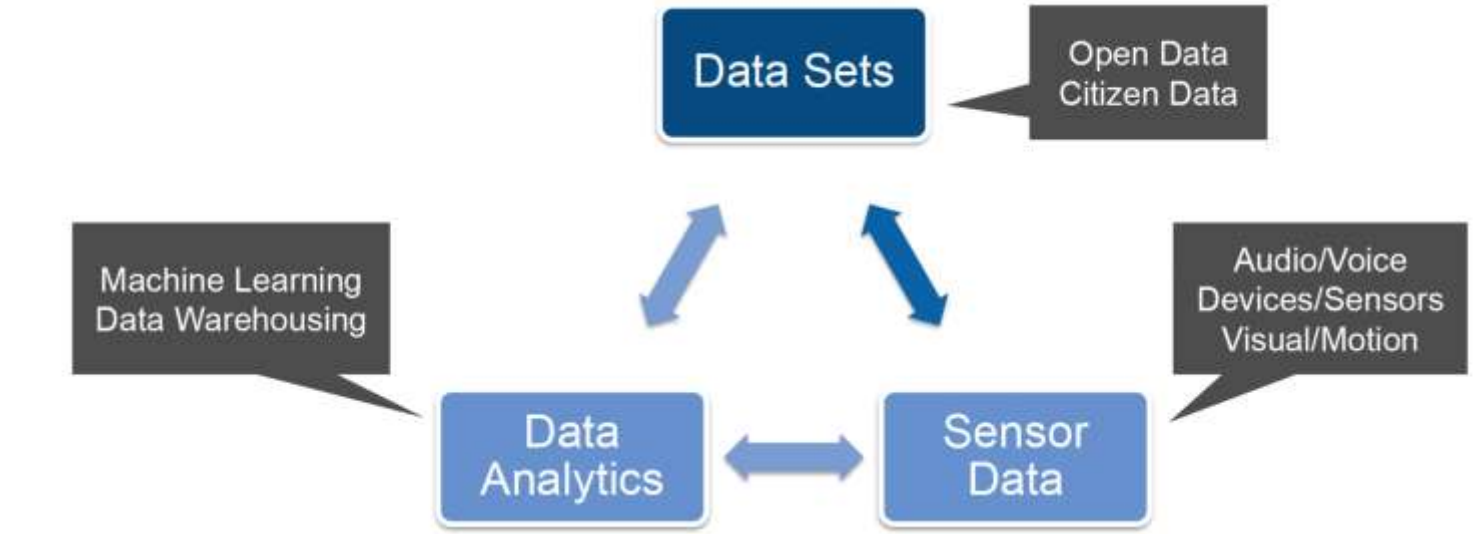


Big Data Problem in IoT

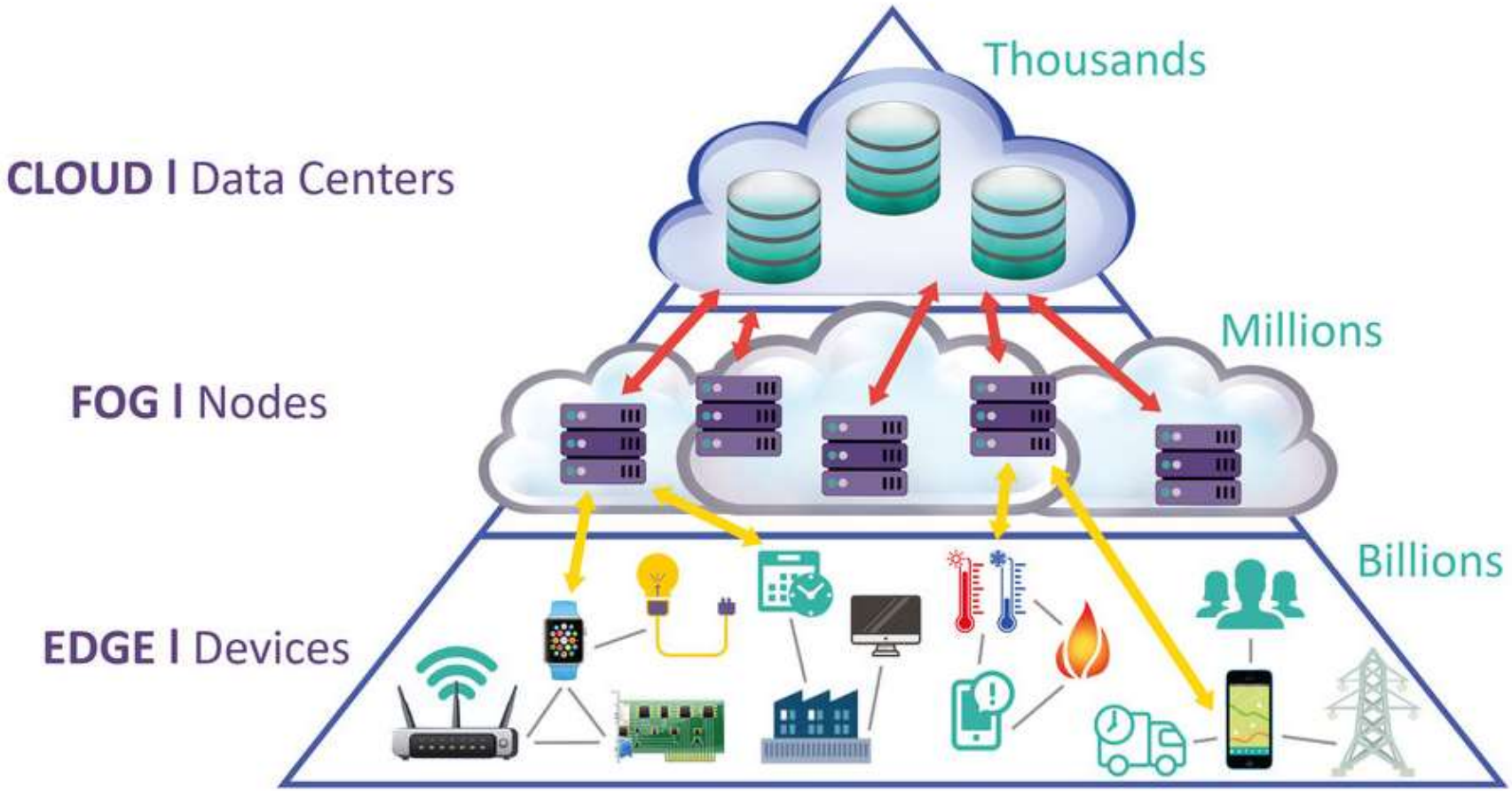
Big Data Problem in IoT (Example)



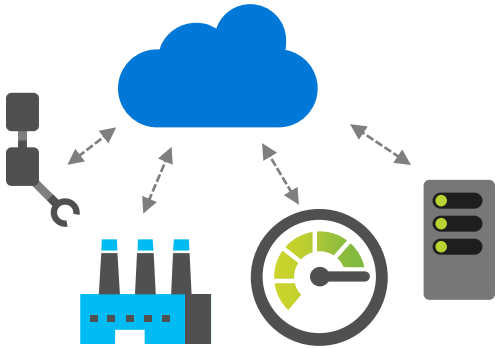
Big Data Processing in IoT

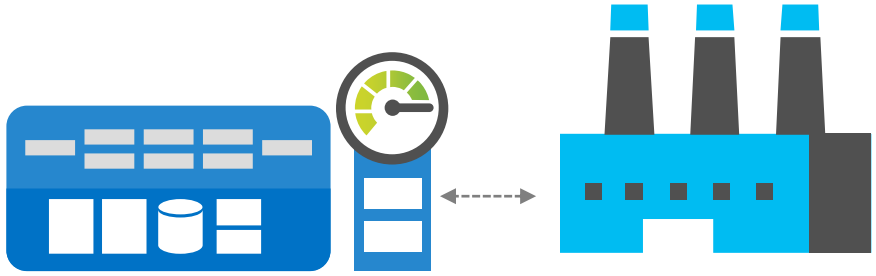


Cloud & Fog Fusion



IoT in the cloud and on the edge





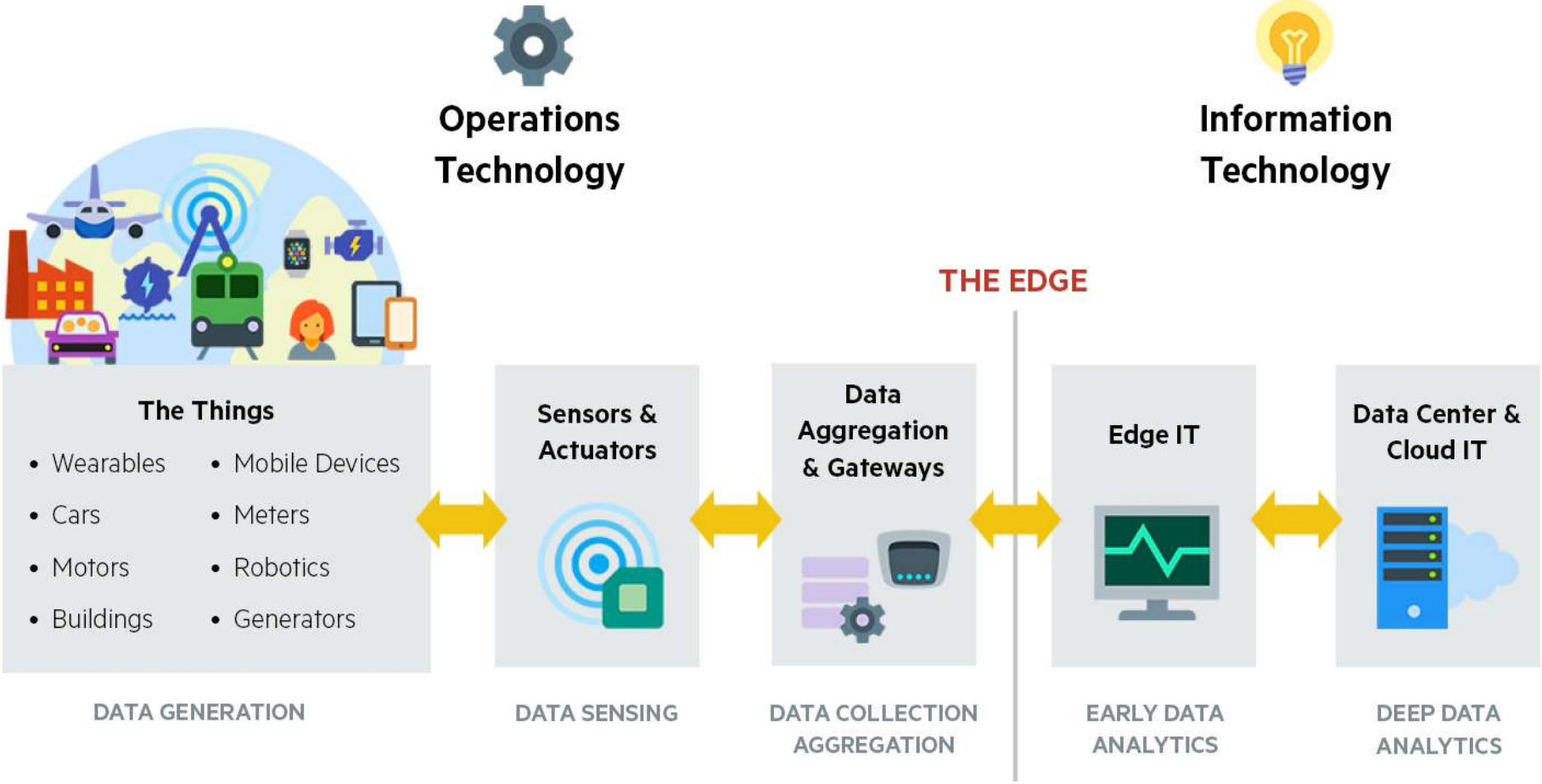
IoT in the Cloud

- 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

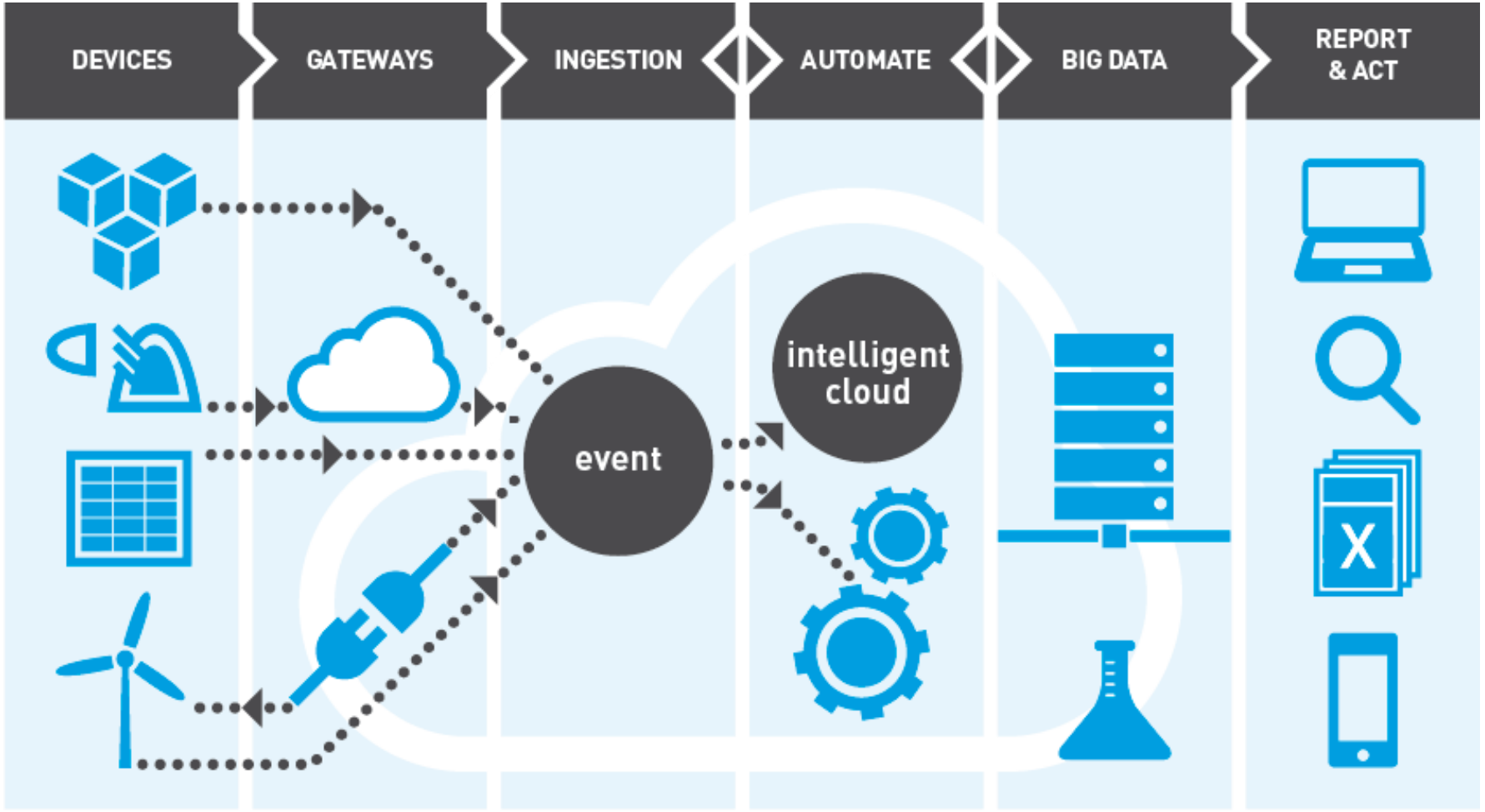
IoT on the Edge

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

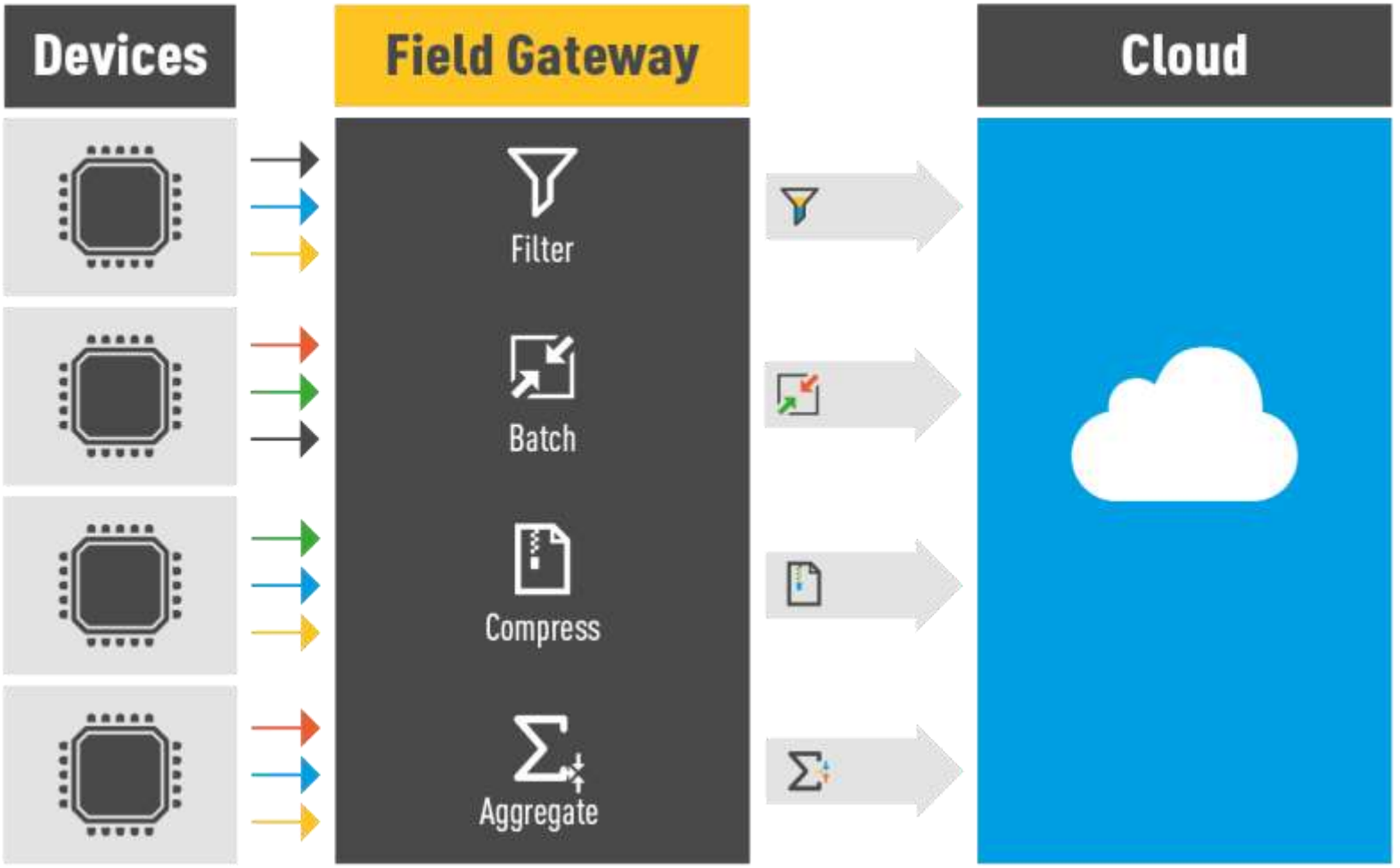
Fog/Edge Computing is the Primary Choice to Handle Real Time Data



IoT End-to-End Value Chain



Smart Gateway



Docker container

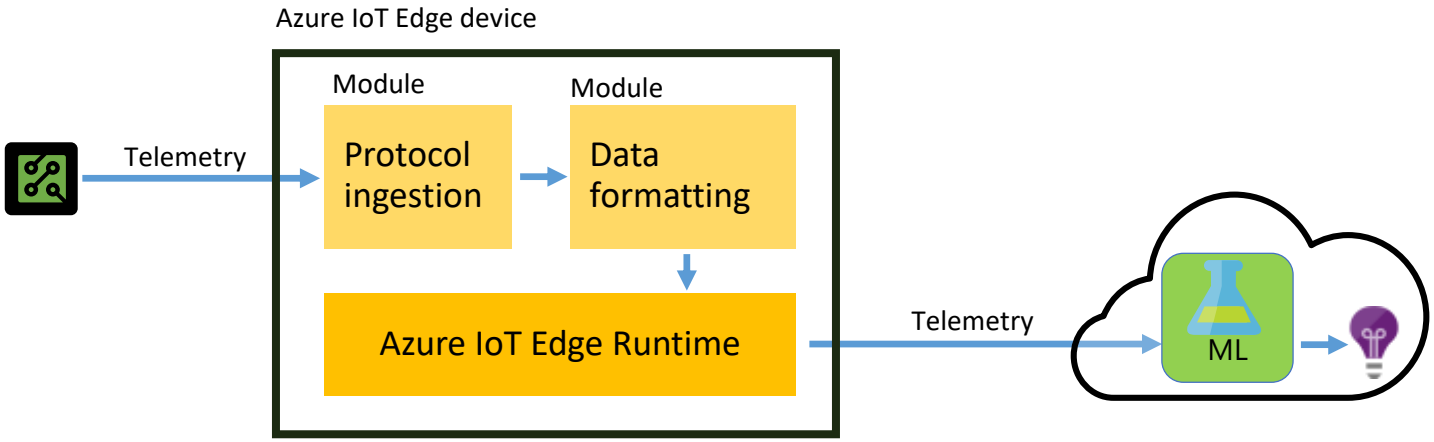
Edge Runtime manages modules

Modules add capabilities to the runtime

Each module performs an action

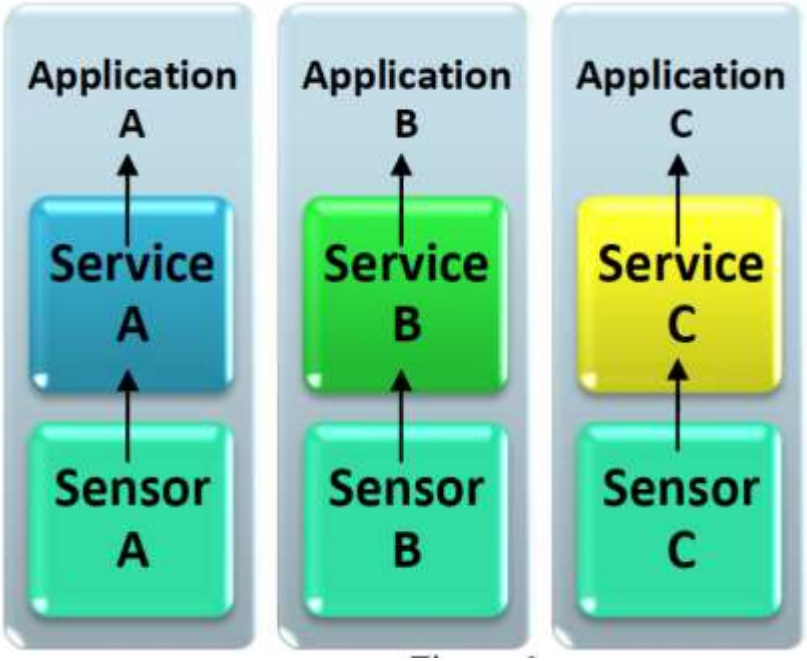
Chain of modules can be thought of as a data processing pipeline, solving an end to end scenario

Modules are Docker containers

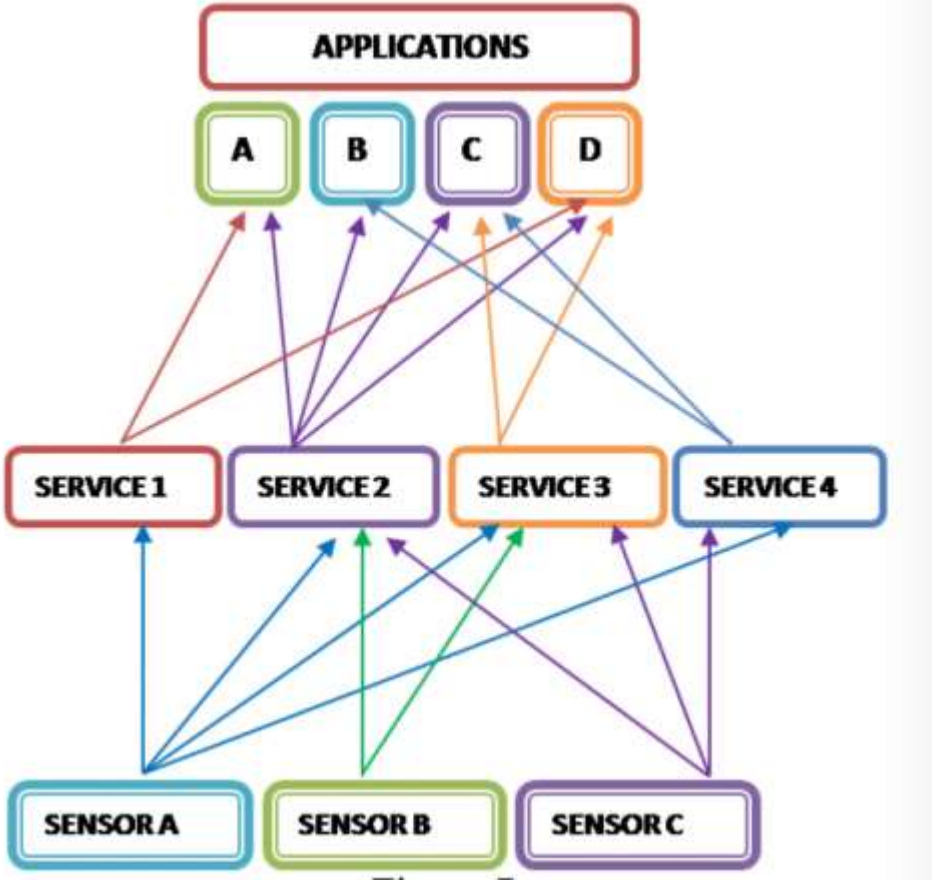


Interoperability problem in IoT

No Interoperation

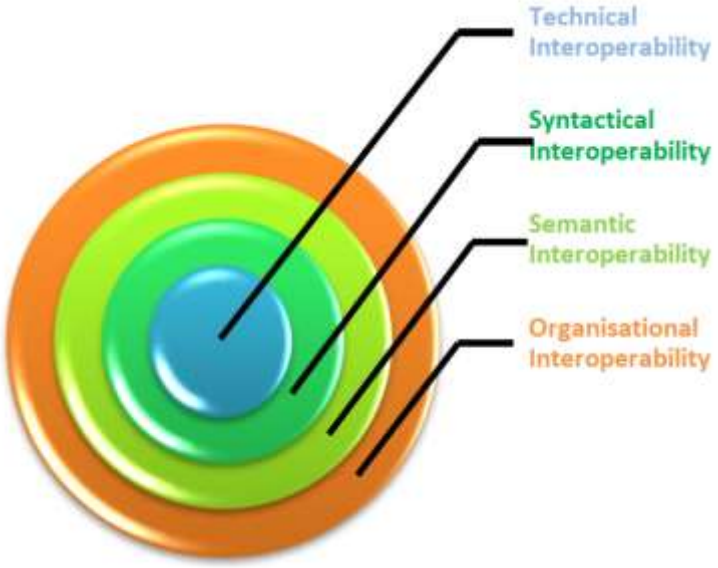


Interoperation

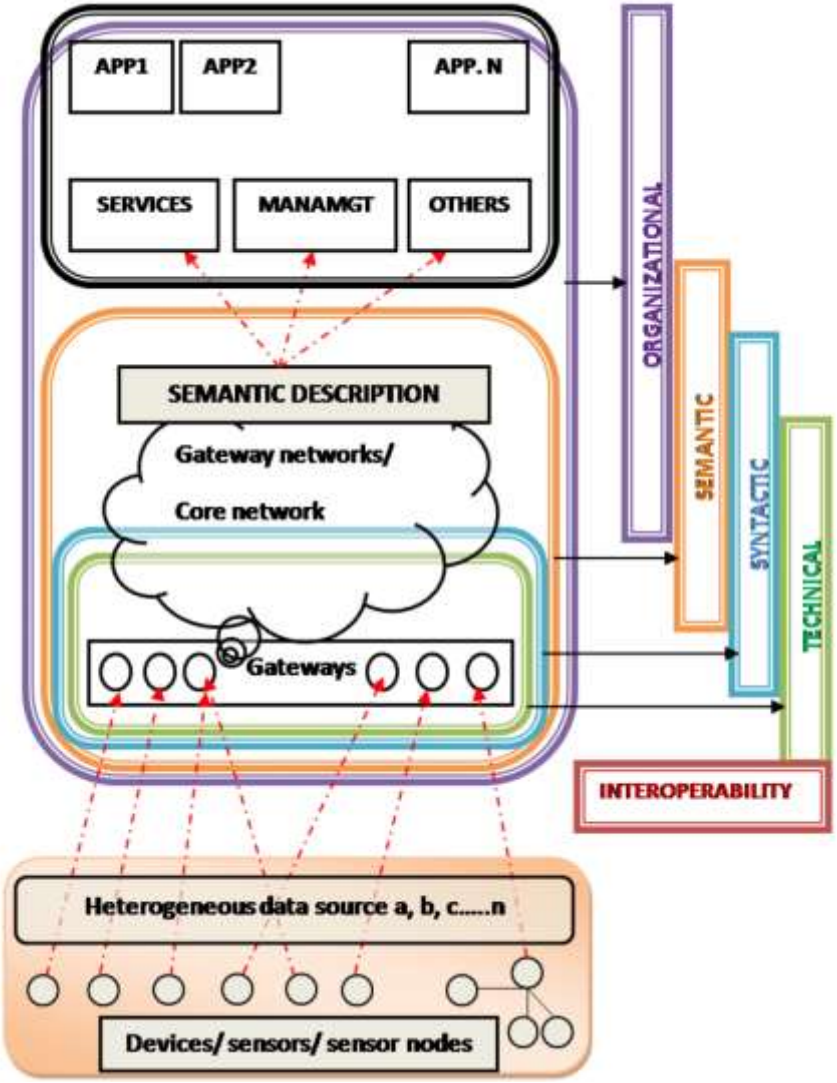


Interoperability solution (1)

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



Interoperability solution (2)



Thank you!