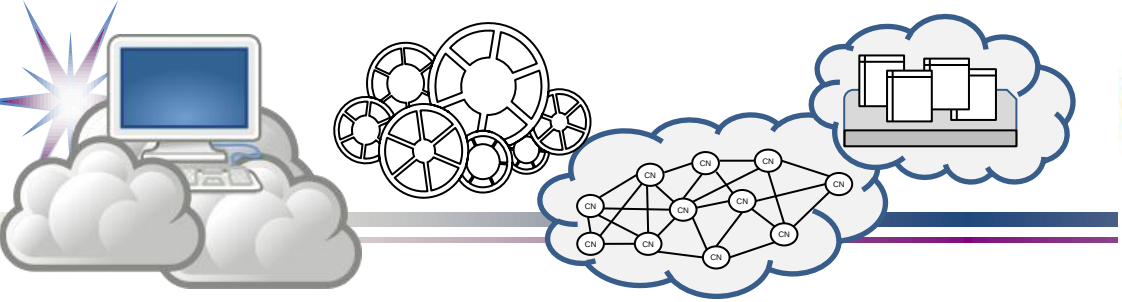


6 листопада 2013, 14:30 Семінар на кафедрі Системного Проектування, КПІ

Сучасні Хмарні технології та технології Великих Даних: Можливості, тенденції та виклики для науки та освіти

Юрій Демченко - кандидат технічних наук, старший науковий співробітник Університету Амстердама (University of Amsterdam)

Теми презентації та питання для дискусії

1. Характеристики і можливості хмарних технологій, тенденції розвитку та стандартизація.

2. Приклади використання та типи впровадження комп'ютерних хмар: корпоративні, публічні, комунальні; міграція корпоративної IT інфраструктури на хмарну платформу, необхідні передумови і рівень "зрілості", переваги віртуалізації сервісів і ресурсів.

3. Законодавча та регуляторна база в Європі, програми підтримки впровадження хмар в Європі.

4. Глобальні провайдери хмарних послуг і ресурсів: Amazon AWS, Microsoft Azure, GoogleCloud: можливості, послуги, засоби розробки.

5. Великі Дані: Об'єм, Швидкість, Номенклатура, Мінливість, Цінність , Достовірність (Volume, Velocity, Variety, Variability, Value, Veracity).

6. Великі Дані та бізнес- аналітика: приклади використання і нові можливості.

7. Проблеми Великих Даних: зберігання, передача, обробка, контроль доступу, захист даних і персональної інформації.

8. Нові спеціальності для Хмарних технологій та Великих Даних: підготовка фахівців, тренінг та освіта.

Сучасні Хмарні технології та технології Великих Даних: Можливості, тенденції та виклики для бізнесу та науки

Cloud and Big Data Technologies: Opportunities and Challenges for business and science

Yuri Demchenko
SNE Group, University of Amsterdam

Дискусія в Часописі,
5 листопада 2013, Київ

http://www.uazone.org/demch/presentations/kiev2013chaspys-cloud-bigdata-biz-v02.pdf

Outline

- Cloud Computing definition and features
- Cloud Computing Reference Architecture and standardisation
- Amazon AWS IaaS cloud and ***cloud powered design*** principles
- European Cloud Computing strategy and legislation

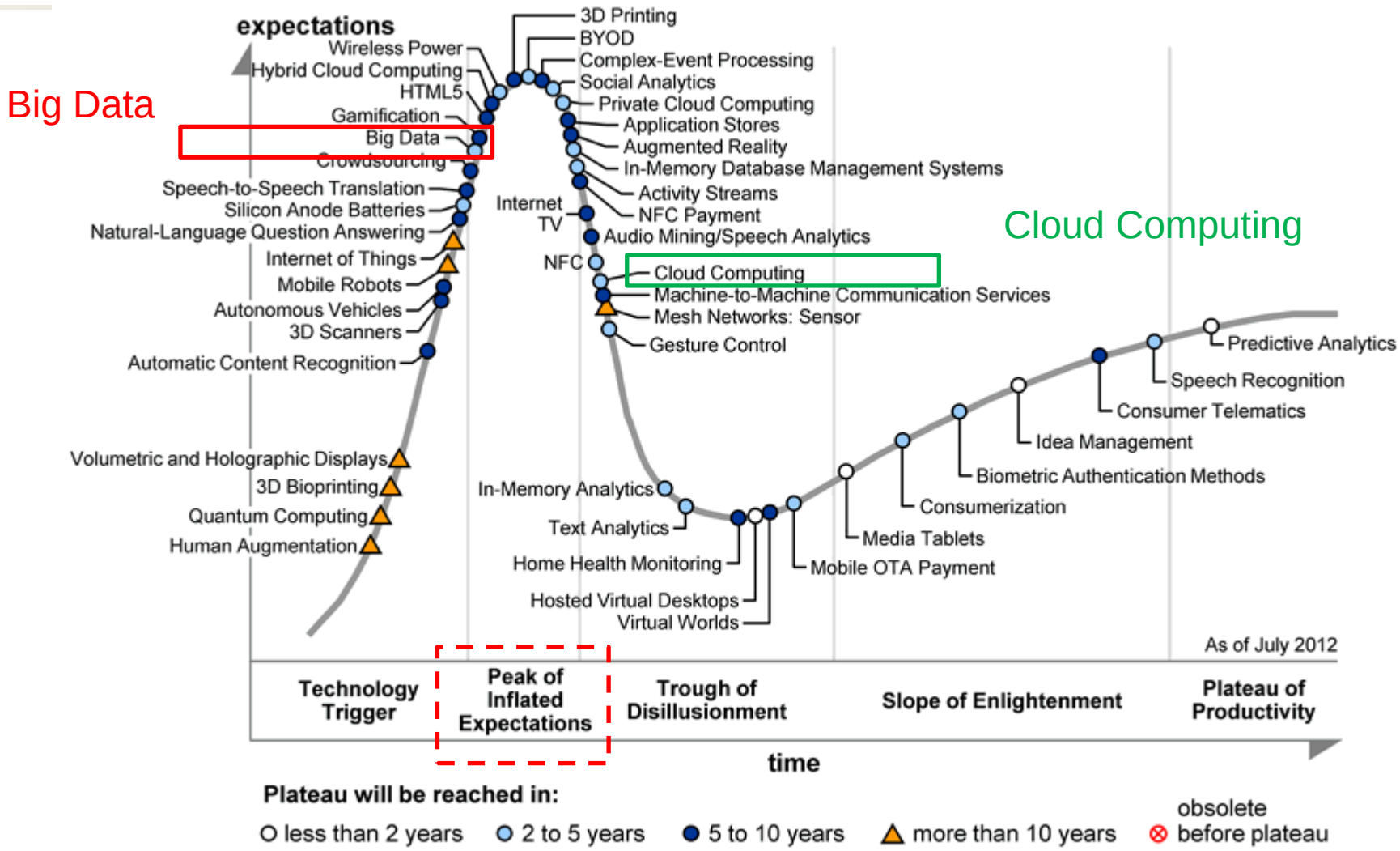
- Big Data and Data Intensive Science as a new technology wave
  - Big Data 5+1 Vs: Volume, Velocity, Variety, Value, Variability, Veracity
- Big Data in Science, Industry and Business
  - Where do the data come from? What are Big Data drivers?
- Defining Big Data Architecture Framework (BDAF)
  - Big Data Infrastructure (BDI) and Big Data Analytics tools

- Data Scientist: New profession and need for Education&Training
- Summary and Discussion

Big Data Research at System and Network Engineering (SNE), University of Amsterdam

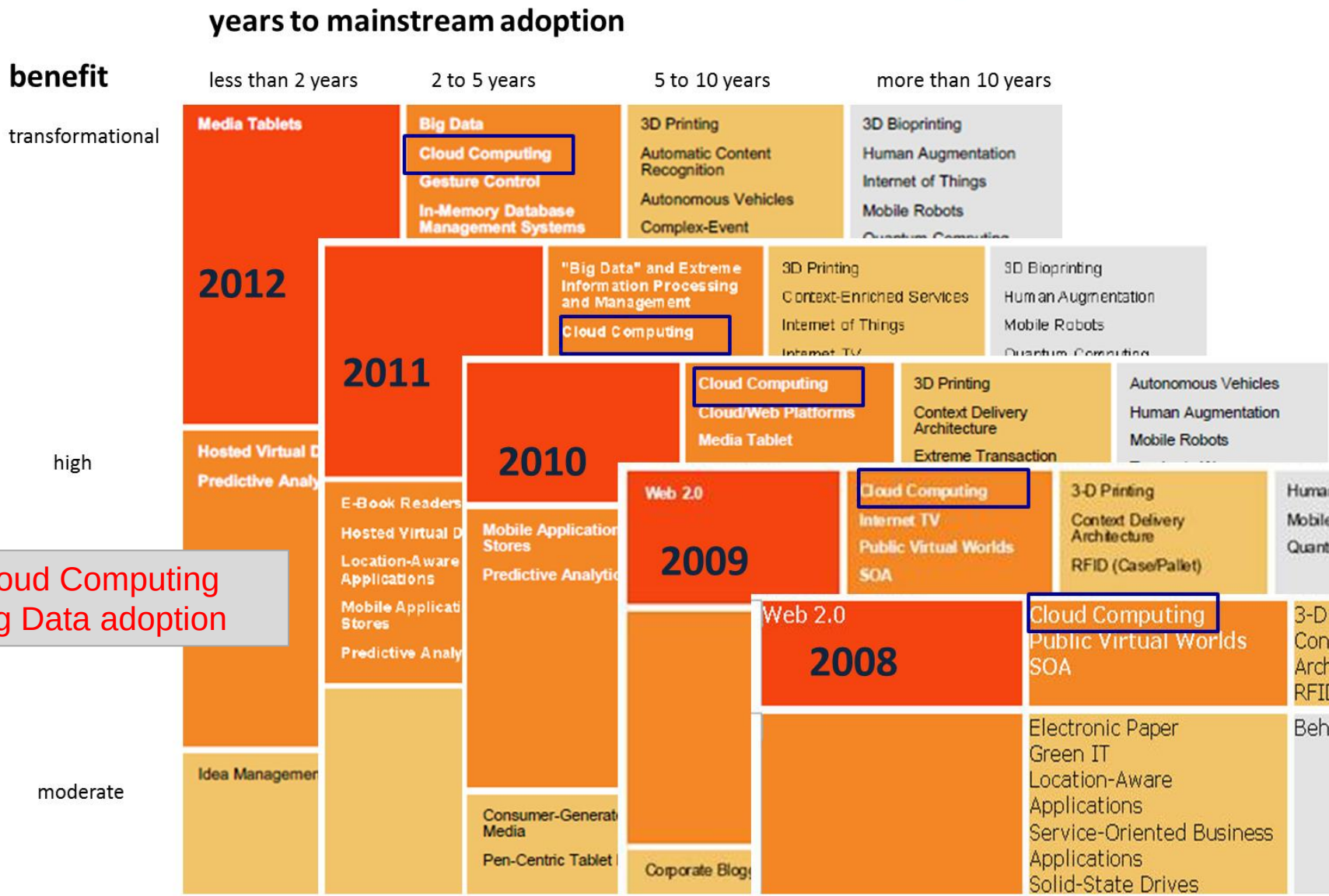
- Long time research and development on Infrastructure services
  - High speed optical networking and data intensive applications
  - Application and infrastructure security services
  - Collaborative systems, Grid, Clouds and currently Big Data
- Focus on Infrastructure definition and services
  - Software Defined Infrastructure based on Cloud/Intercloud technologies
- Standardisation activity (IETF, OGF, ISO - past)
  - NIST Cloud Computing Reference Architecture (past) and Big Data Working Group
  - Research Data Alliance: Education and Data Analytics Tools
- Big Data Interest Group at SNE, UvA
  - Non-formal but active, meets two-weekly for brainstorming sessions
  - Provides input to NIST BD-WG and RDA
- SNE Technical Report on Architecture Framework and Components for the Big Data Ecosystem. Draft Version 0.2, 12 September 2013
  http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

Gartner Technology Hypercycle



Source http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Gartner Priority Matrix



5 yr for Cloud Computing
2 yr for Big Data adoption

Technology Definitions and Timeline - Overview

- **Service Oriented Architecture (SOA)**: First proposed in 1996 and revived with the Web Services advent in 2001-2002
  - Currently standard for industry, and widely used
  - Provided a conceptual basis for Web Services development
- **Computer Grids**: Initially proposed in 1998 and finally shaped in 2003 with the Open Grid Services Architecture (OGSA) by Open Grid Forum (OGF)
  - Currently remains as a collaborative environment
  - Migrates to cloud and inter-cloud platform
- **Cloud Computing**: Initially proposed in 2008 – *Now entering productive phase*
  - Defined ***new features, capabilities, operational/usage models*** and actually provided a guidance for the new technology development
  - Originated from the Service Computing domain and service management focused
- **Big Data and Data Intensive Science**: *Yet to be defined*
  - Involves more components and processes to be included into the definition
  - Can be better defined as Ecosystem where data are the main driving component
  - Need to define the Big Data properties, expected technology capabilities and provide a guidance/vision for future technology development

Big Data and Clouds and Mobile Technologies – **From disruptive to consolidating technologies**

- Service Oriented Architecture (SOA) - Industry
- Computer Grids, Distributed Computing – Research community
- Cloud Computing: Initially proposed in 2008 – *Now entering productive phase (Industry)*
  - Functional Cloud Computing definition provided a guidance for the technology development
  - Consolidating SOA, Distributed computing, SDN
  - Facilitated by Mobile technologies, Big Data
- **Big Data and Data Intensive Science** – *Originated from science*
  - Consolidates Cloud Computing, Mobile technologies, High Performance computing, Data warehousing, Data analytics/science
  - Emerges new data centric models and technologies
    - Introduces new technical category Ecosystem where data are the main driving component
  - Need to define the Big Data properties, expected technology capabilities and provide a guidance/vision for future technology development

Основы технологии Облачных Вычислений

- Cloud Computing technology foundation

Анатомия Облачных Технологий

- Обьединние большого количества вычислительных ресурсов и запоминающих устройств в больших Центрах Обработки Даных (ЦОД)
  - Экономия размера/масштаба
  - Горизонтальное и вертикальное масштабирование: балансировагние нагрузки и эластичность
- Виртуализация сервисов, ресурсов и платформ
  - Virtualisation == (Pooling) – Abstraction – Composition – Deployment – (Lifecycle management)
  - Механизмы привязки: пространство имен и безопасность/доверительность
  - Миграция виртуальных машин (ВМ; VM - Virtual Machine)
- Многоуровневая модель виртуализации сервисов и ресурсов
  - Много-пользовательская среда (multi-tenancy), учет, билинг
- Всеобщий и универсальный доступ через Интернет, всеобщая коннективность

Top Cloud Providers (July 2013) – Ranked 1-20

1. Salesforce.com (revenue >$3Bln)
2. Amazon AWS (>$1.5Bln)
3. Microsoft (>$1.5Bln)
4. Oracle
5. Google
6. SAP
7. SoftLayer (IBM since 2013) + IBM's SmartCloud
8. Terremart (Verizon Company)
9. Rackspace
10. NetSuite ERP cloud service provider ($308Mln)
11. Workday HR and financial cloud services
12. Dropbox
13. Savvis (CenturyLink Company) Oracle and private clouds
14. Joyent
15. Navisite (Time Warner) enterprise applications from IBM, MS, Oracle
16. Citrix Systems
17. LogMeIn Remote Management platform provider
18. Zoho – alternative to Salesforce, Offcie365, Google Apps
19. Dimension Data (NTT Group, ZA) managed hosting

20 Carbinite backup provider for SQL, MS Exchange

European Cloud Computing Strategy (2012+)

In September 2012, the European Commission adopted a strategy for ”Unleashing the Potential of Cloud Computing in Europe”

- To deliver 2.5 mln new European jobs, and an annual boost of EUR 160 billion to EU GDP (around 1%)

Key actions

- Safe and Fair Contract Terms and Conditions
  - **data preservation after termination of the contract**
  - data disclosure and integrity
  - data location and transfer
  - ownership of the data
  - direct and indirect liability change of service by cloud providers and subcontracting
- Cutting through the Jungle of Standards
  - EU cloud framework to emerge in next 18+ months, by 2015
- Establishing a European Cloud Partnership (ECP)

- Mixed reactions from the major cloud players in Europe (e.g., Microsoft, HP, IBM, majority from US)

NIST Cloud definition – NIST SP 800-145 (1)

NIST SP 800-145 The NIST Definition of Cloud Computing (Draft)
http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf

- Five Cloud characteristics
  - On-demand self-service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured Service
- 3 basic service models
  - Software as a Service (SaaS)
  - *Platform as a Service (PaaS)*
  - Infrastructure as a Service (IaaS)
- Deployment models
  - Private clouds
  - Public clouds
  - Hybrid clouds
  - Community clouds

NIST Cloud definition – NIST SP 800-145 (1)

NIST SP 800-145 The NIST Definition of Cloud Computing (Draft)
http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf

- Пять основных характеристик Облачных технологий
  - Самообслуживание по требованию
  - Широкополосный сетевой доступ
  - Агрегатирование (пулинг) ресурсов
  - Быстрая эластичность
  - Измеряемые услуги (учет обьема услуг)
- 3 базовые модели услуг
  - Software as a Service (SaaS)
  - *Platform as a Service (PaaS)*
  - Infrastructure as a Service (IaaS)
- Модели реализации
  - Private clouds
  - Public clouds
  - Hybrid clouds
  - Community clouds

NIST Cloud definition – Draft SP 800-145 (2)

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., *networks*, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.

Cloud Infrastructure as a Service (IaaS)

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, deployed applications, and *possibly limited control of select networking components (e.g., host firewalls)*.

Note: NIST Definition of Cloud – missing network provisioning, just “limited control over network”.

NIST Cloud Definition – PaaS, SaaS

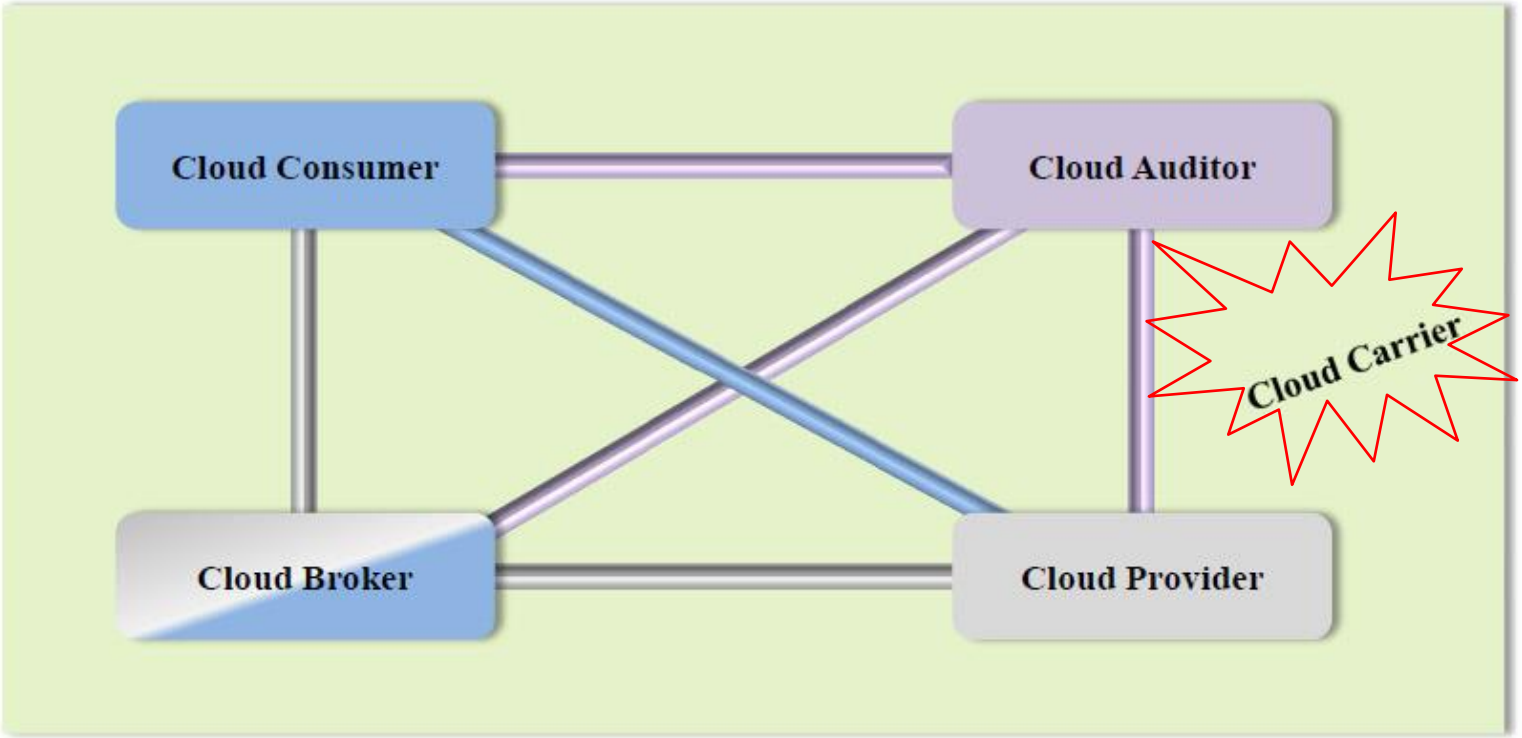
Platform as a Service (PaaS)

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*.

Software as a Service (SaaS)

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*.

NIST Cloud Computing Reference Architecture (CCRA) 2.0 - Main Roles (1)



The communication path between a cloud provider and a cloud consumer

The communication paths for a cloud auditor to collect auditing information

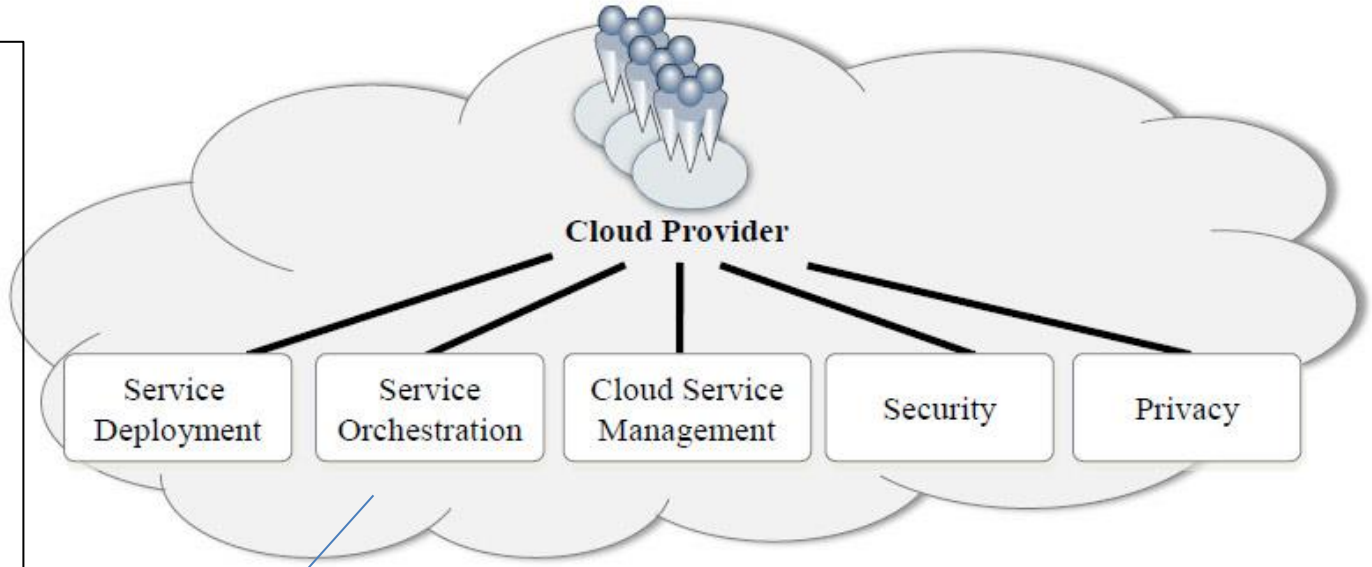
The communication paths for a cloud broker to provide service to a cloud consumer

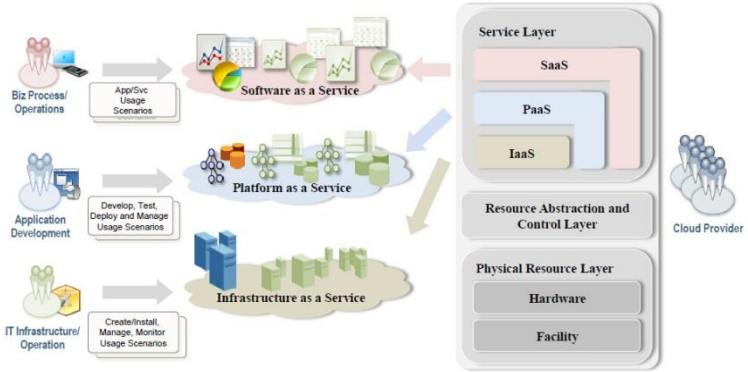
NIST Cloud Computing Reference Architecture (CCRA) 2.0 - Provider Functions (2)

Service provisioning stages:

Service Delivery (Framework)

- Request&SLA
- Reservation/ Composition
- Deployment
- Operation
- Decommissioning

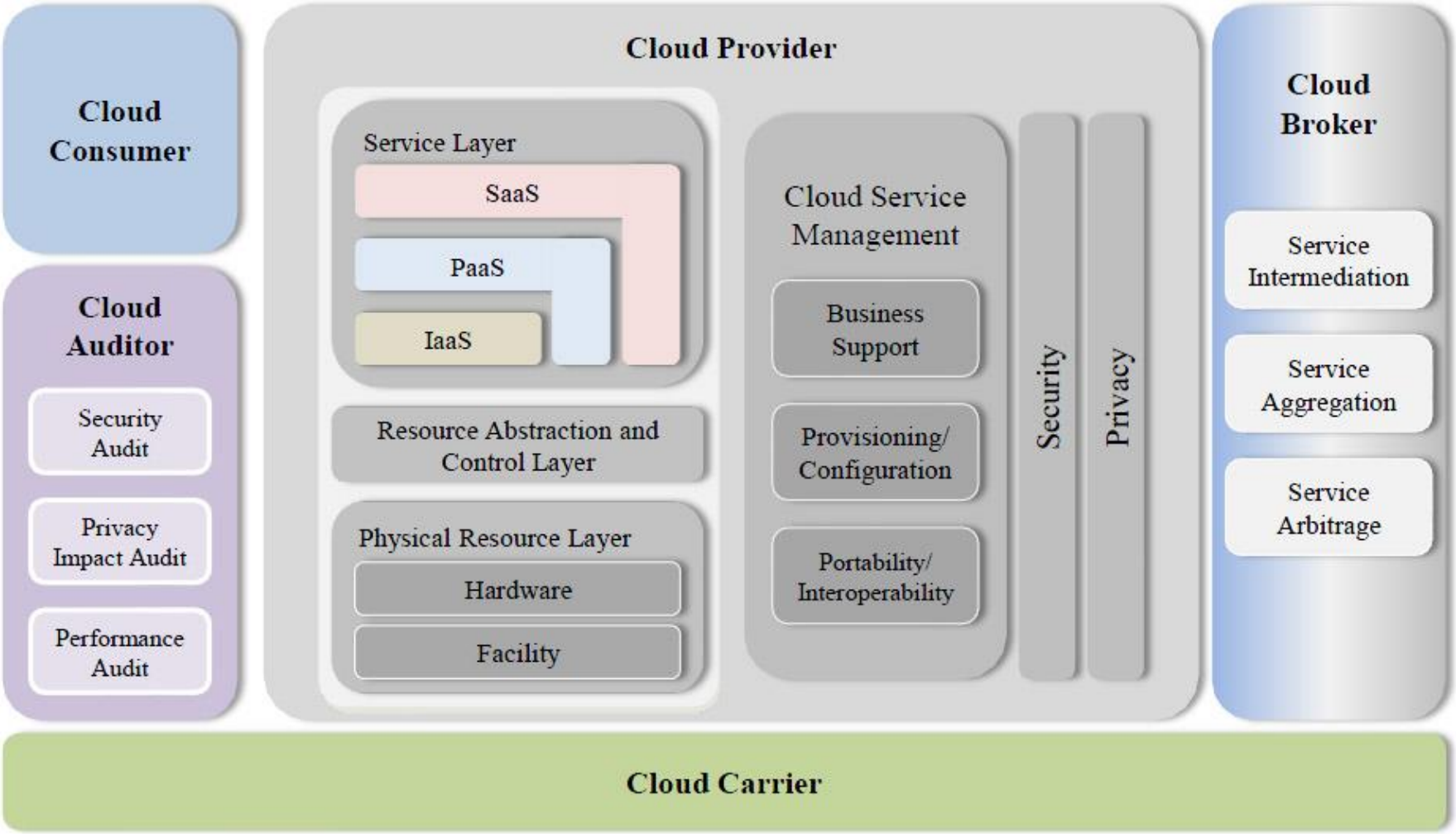




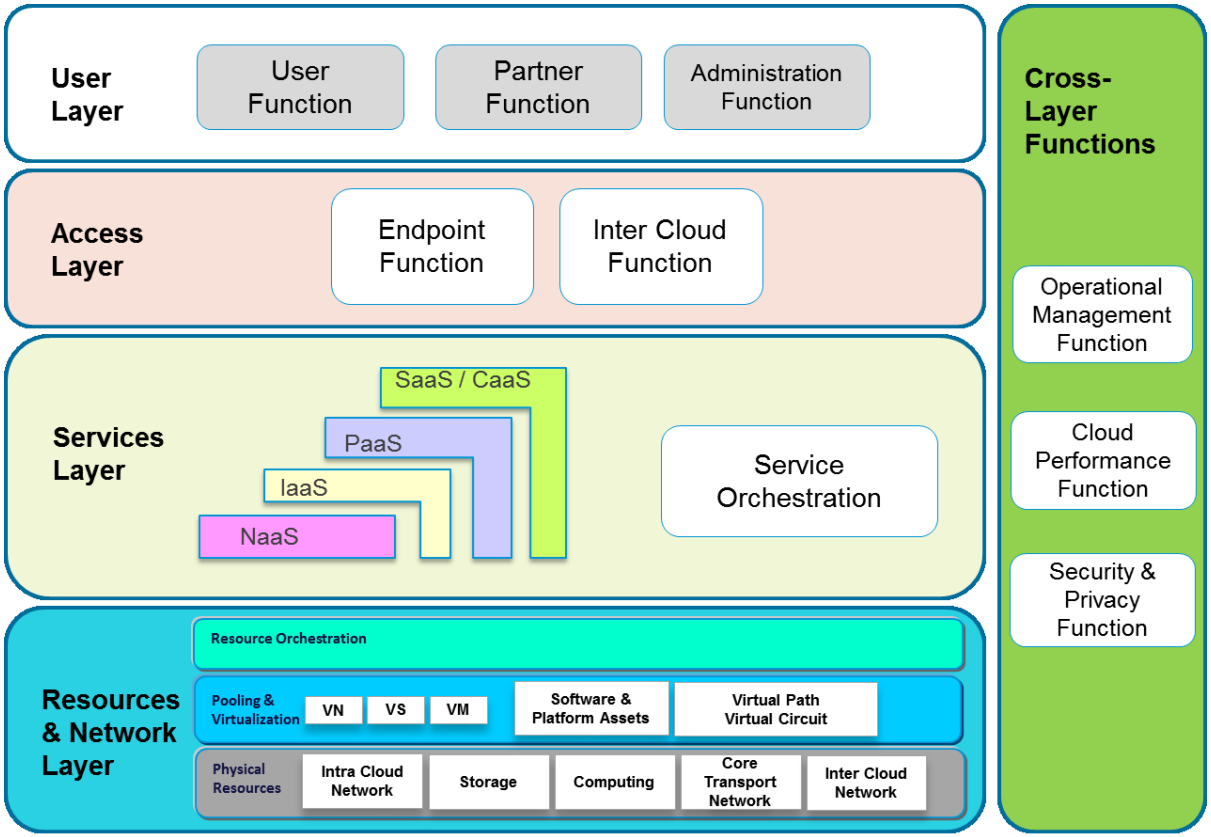
Функции провайдера облачных услуг

- Внедрение.установка сервисов и ресурсов
- Взаимодействие сервисов и ресурсов
- Управление ресурсами и сервисами
- Безопасность
- Приватность

NIST Cloud Computing Reference Architecture (CCRA) 2.0 – Consolidated View (3)



ITU-T FG on Clouds Part 2: Functional requirements and reference architecture



Layered Cloud computing architecture includes:

- User layer (including user functions, partner functions, administration functions)
- Access layer (including endpoint functions and inter-cloud functions). Network service providers role is to provide inter-cloud transport network
- Cloud services layer (including basic cloud services IaaS, PaaS, SaaS, NaaS, CaaS and also Orchestration service)
- Resources and network layer (including physical resources, pooling and orchestration, pooling and virtualisation)

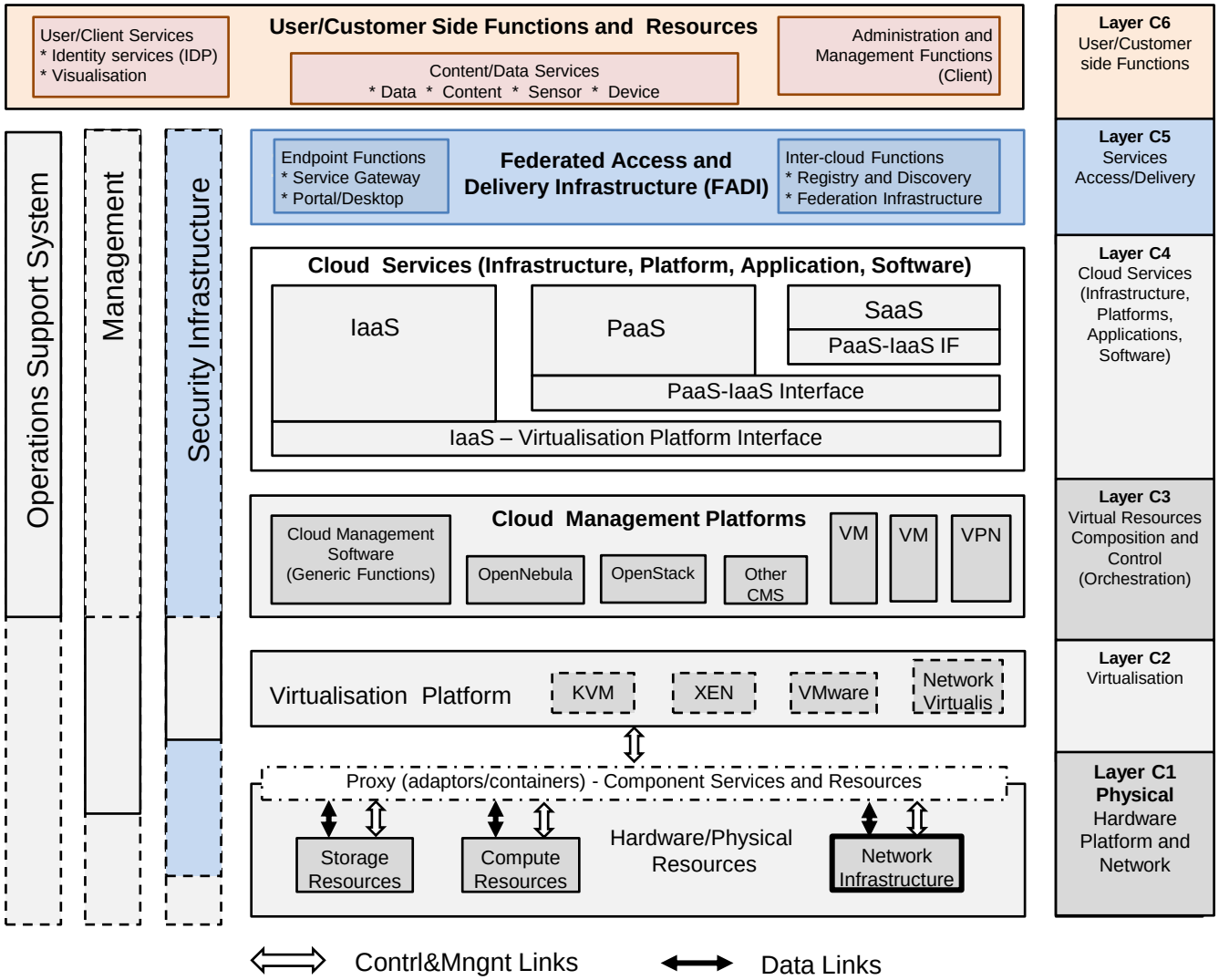
Cloud Reference Framework (IETF I-Draft)

Cloud Reference Framework. Internet Draft, by B. Khasnabish, J. Chu, S. Ma, Y. Meng, N. So, P. Unbehagen, M. Morrow, M. Hasan, Y. Demchenko
http://tools.ietf.org/html/draft-khasnabish-cloud-reference-framework-05.txt

- Multilayer Cloud Services Model (CSM)
  - Including Federated Access and Delivery Infrastructure layer and user side services layer
- Intercloud Architecture Framework (ICAF)
  - InterCloud Control and Management Plane (ICCMP)
    - Signaling, monitoring, synchronisation between heterogeneous clouds
  - InterCloud Federation Framework (ICFF)
    - Protocols and mechanisms for heterogeneous clouds integration
  - InterCloud Operations and Management Framework (ICOMF)
    - Services and business processes management and operation
  - Intercloud Security Framework (ICSF)

Multilayer Cloud Services Model (CSM)

http://www.ietf.org/id/draft-khasnabish-cloud-reference-framework-05.txt



CSM layers

(C6) User/Customer side Functions

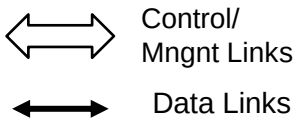
(C5) Intercloud Access and Delivery Infrastructure

(C4) Cloud Services (Infrastructure, Platform, Applications)

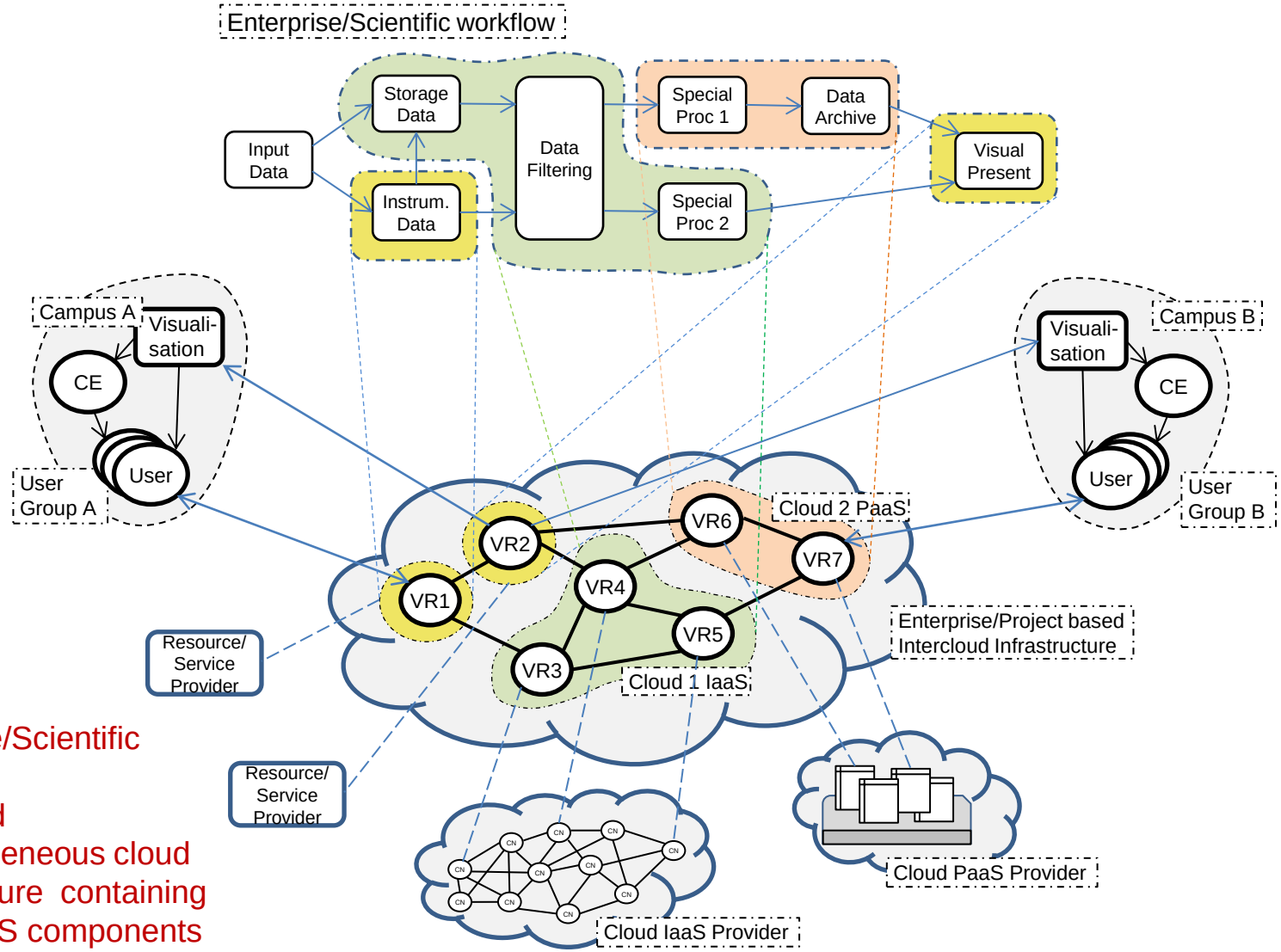
(C3) Virtual Resources Composition and Orchestration

(C2) Virtualisation Layer

(C1) Hardware platform and dedicated network infrastructure



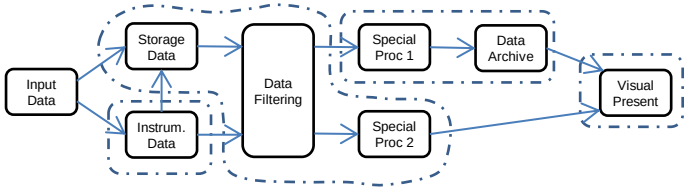
General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



Enterprise/Scientific workflow
Is mapped
to heterogeneous cloud infrastructure containing IaaS, PaaS components

Intercloud Applications Integration: ICCM, ICFF, ICOMF

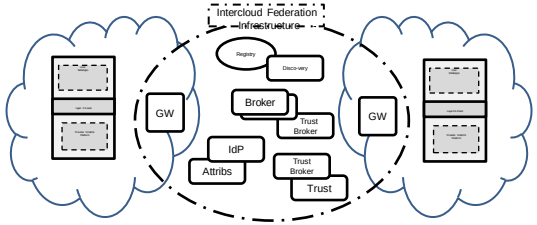
ICOMF – InterCloud Operations and Mngnt Framework

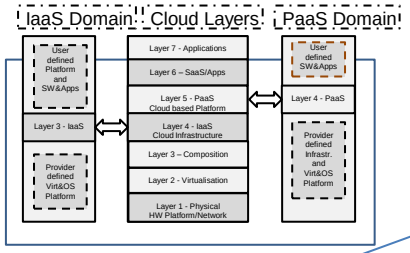


Business Processes Management and Services Operation Support

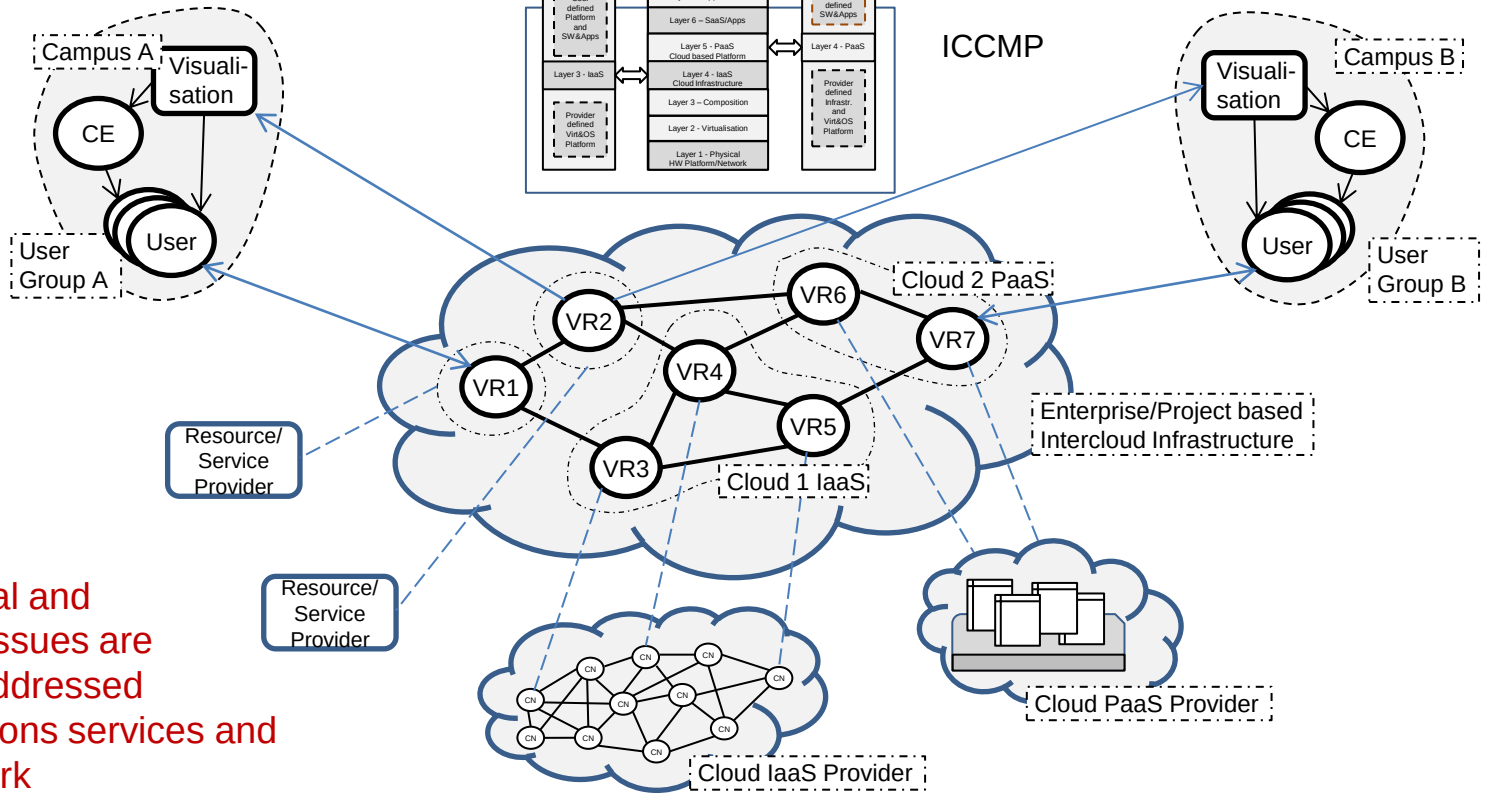
- SLA Management
- Business roles and Actors
- Business level Service Registry and Broker
- Mobility?

ICFF – InterCloud Federation Framework





ICCMP



Operational and business issues are typically addressed by Operations services and a framework

Cloud IaaS – облачные инфраструктурные сервисы

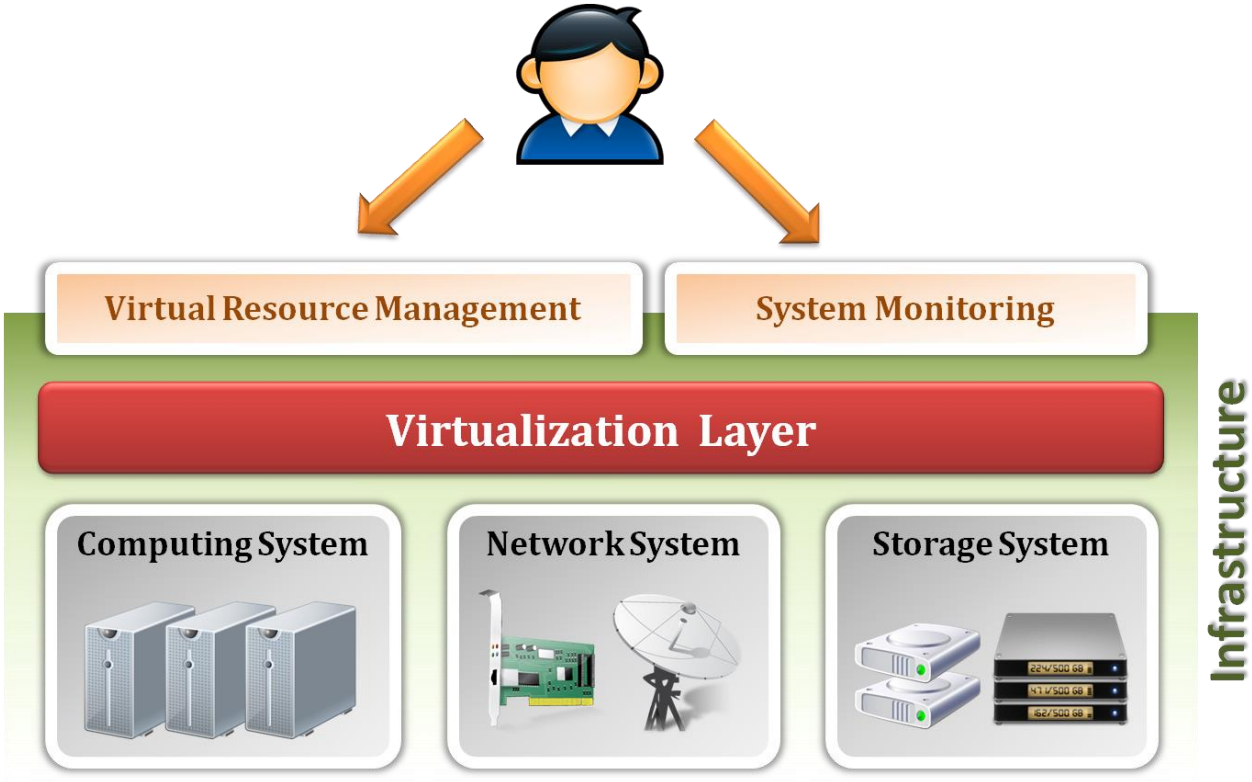
Cloud IaaS – Основные свойства и характеристики облачных систем и инфраструктур

- Все фундаментальные свойства и характеристики компьютерных облаков должны быть адресованы и внедреныв IaaS

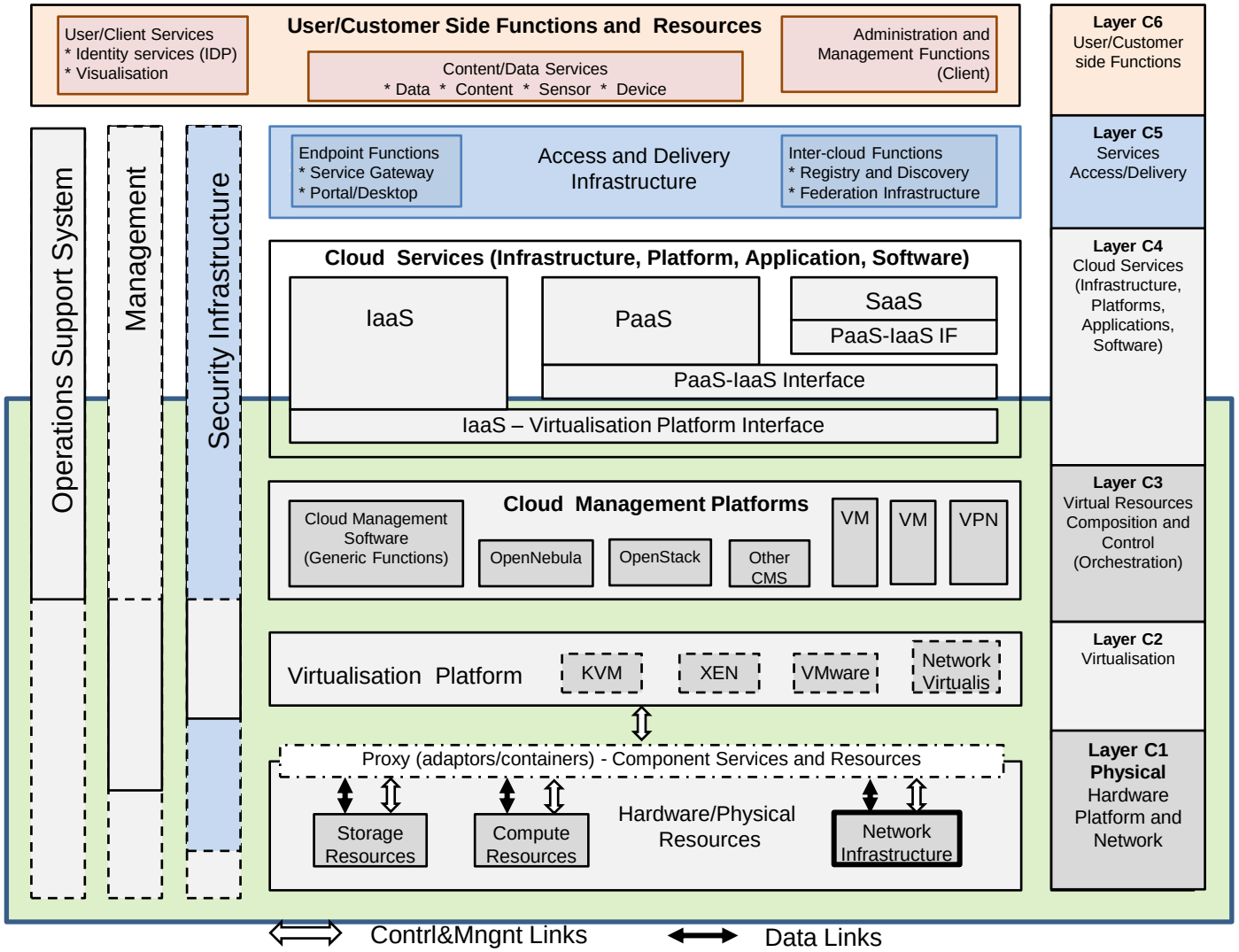
- Масштабируемость и эластичность
  - Динамичное внедрение и установка
  - Много-пользовательская среда
- Доступность и надежность
  - Устойчивость к сбоям
  - Устойчивость системы
  - Безопасность
- Управляемость и совместимость
  - Управляющая информация
  - Мониторинг систем
  - Биллинг
- Производительность и оптимизация
  - Параллельная обработка
  - Баланс нагрузки
  - Планирование работ
- Доступ и переносимость
  - Стандартный доступ (через Интернет)
  - Тонкий клиент
  - Стандартизация

Cloud IaaS Architecture

- **Infrastructure as a Service (IaaS)** delivers computer infrastructure for cloud user, typically a platform virtualization environment as a service.
- **Virtualization** is an enabling technique to provide an abstraction of logical resources away from underlying physical resources.



Multilayer Cloud Services Model (CSM)



CSM layers

(C6) User/Customer side Functions

(C5) Intercloud Access and Delivery Infrastructure

(C4) Cloud Services (Infrastructure, Platform, Applications)

(C3) Virtual Resources Composition and Orchestration

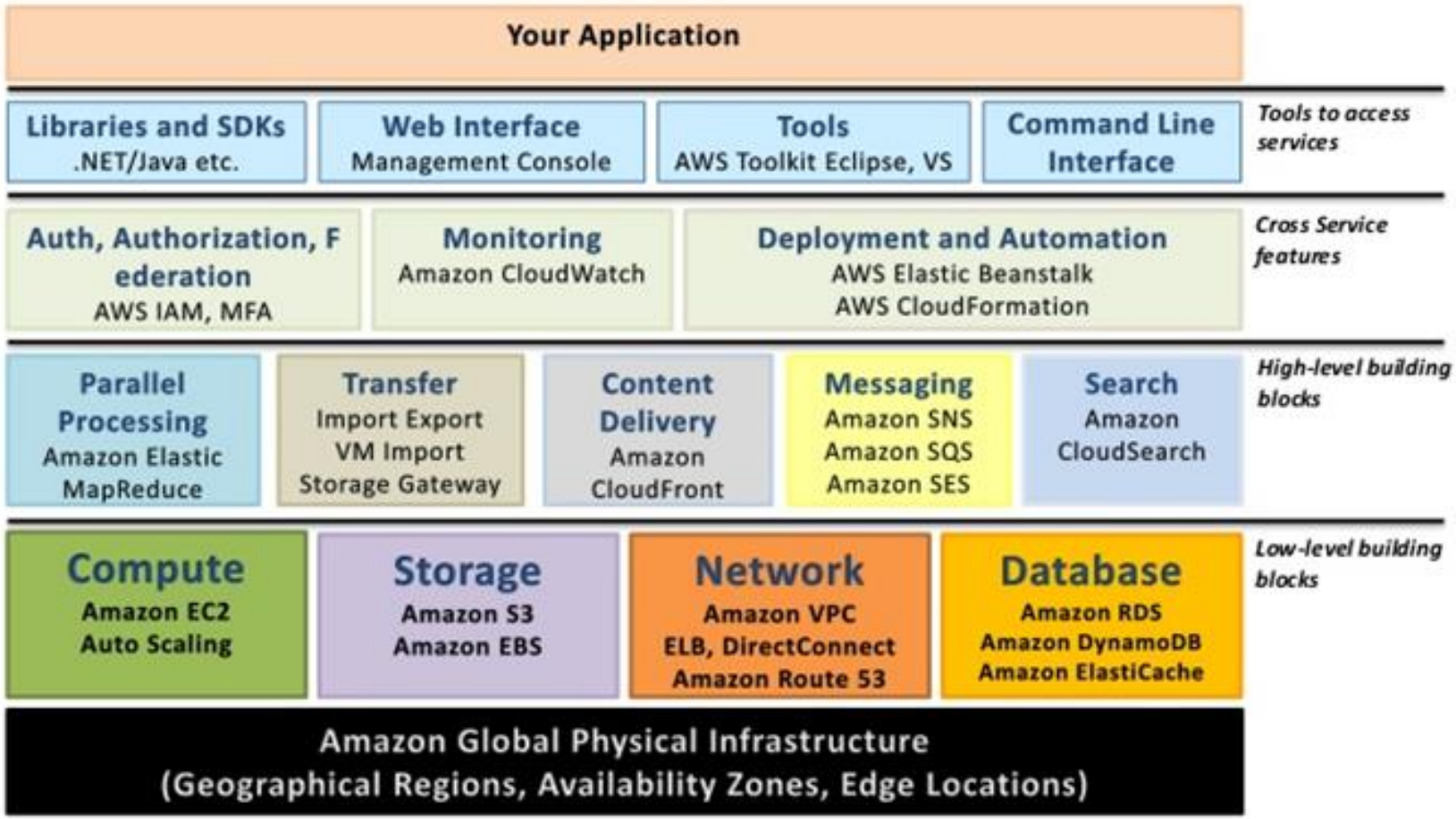
(C2) Virtualisation Layer

(C1) Hardware platform and dedicated network infrastructure

Control/ Mngnt Links

Data Links

Amazon AWS Cloud Architecture



Credits “Building Powerful Web Applications in the AWS Cloud” by Louis Columbus
http://softwarestrategiesblog.com/2011/03/10/building-powerful-web-applications-in-the-aws-cloud/

Amazon EC2 User Application Component Services

- *EBS Elastic Block Store*
- *Elastic IP Address dynamically assigned to user VMs*
- *VPC (Virtual Private Cloud)* allows organizations to use AWS resources along with their existing infrastructure in a VPN (Virtual Private Network) to increase compute capabilities beyond the local resources.
- *CloudWatch* monitoring service
- *Auto Scaling* dynamic resource provision
- *Elastic Load Balancing* between multiple VMs located within a single availability zone or multiple zones
- *VM Import/Export* for custom VM images store and load
- Cloud Formation services composition and deployment

Amazon S3 (Simple Storage Service)

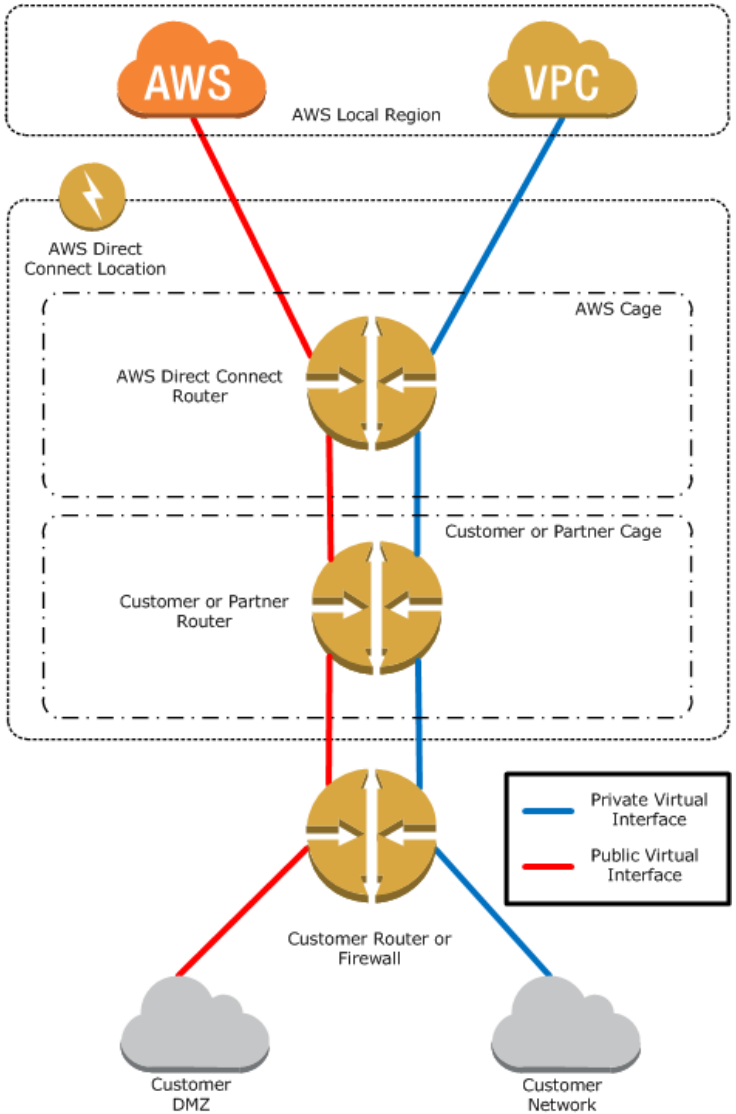
S3 is a service that stores large amounts of data, and is accessible via the Internet

- Amazon S3 is intentionally built with a minimal feature set
- Data stored as objects associated with unique keys
- Objects can store ***up to 5 TB*** of data and are bound to ***specific buckets*** which can only be stored in a particular region (availability zone)

"Data stored in Amazon S3 is *secure* by default" :-); only bucket and object owners have access to the Amazon S3 resources they create

- Amazon S3 supports multiple access control mechanisms, as well as encryption for both secure transit and secure storage on disk
- With Amazon S3's data protection features, you can protect your data from both logical and physical failures, guarding against data loss from unintended user actions, application errors, and infrastructure failures
- For customers who must *comply with regulatory standards such as PCI and HIPAA*, Amazon S3's data protection features can be used as part of an overall strategy to achieve compliance
- Cost issues: free upload, paid download

New Service: Amazon Direct Connect



- AWS Direct Connect links customer internal network to an AWS Direct Connect location over a standard 1 Gbps or 10 Gbps Ethernet fiber-optic cable
  - One end of the cable is connected to customer router, the other to an AWS Direct Connect router
  - Allows creating virtual interfaces directly to the AWS cloud (Amazon EC2, S3) and to Amazon Virtual Private Cloud (Amazon VPC), bypassing Internet service providers in your network path
- Access is limited to Amazon Web Services in the region

Example: Cloud powered services design with AWS

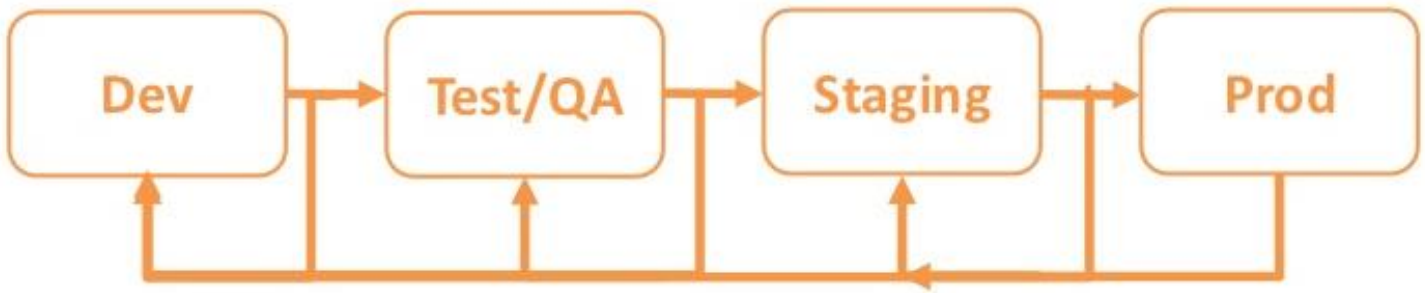
- Based on AWS seminar presentation
  "Building Powerful Web Applications in the AWS Cloud" by Louis Columbus
  http://softwarestrategiesblog.com/2011/03/10/building-powerful-web-applications-in-the-aws-cloud/

  Pattern #1: Design for failure and nothing will fail
  Pattern #2: Edge cache static content
  Pattern #3: Implement Elasticity
  Pattern #4: Leverage Multiple Availability Zones
  Pattern #5: Isolate read and write traffic; Isolate static and dynamic traffic
  Pattern #6: Hardening security at every stage
  Pattern #7: Parallel Processing
  Pattern #8: Go global quickly (with single API)
  Pattern #9: Automate your in-cloud Development and Deployment Lifecycle
  Pattern #10: Keep optimizing and see the savings in the next month's bill

Hardening Security

Credits “Building Powerful Web Applications in the AWS Cloud” by Louis Columbus
http://softwarestrategiesblog.com/2011/03/10/building-powerful-web-applications-in-the-aws-cloud/

Cloud powered development lifecycle



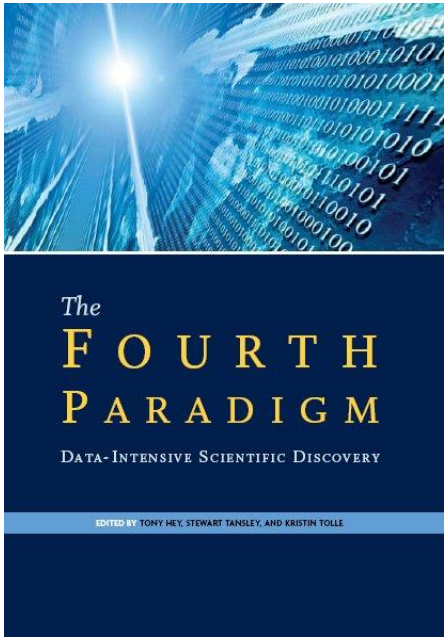
For the Test/QA stage:
Speed: quickly get on-demand resources
Variety: test more demo configurations
Real world load and stress testing; easy simulates 100s of clients
Repeatability: preconfigured shareable Test DB in minutes
Reproducibility: “Save As” Productions Environment and re-launch in Test Environment
Savings: “Turn off” Testing Environment

- Environmental separation
- Environmental consistency
- Variable resource
- Different control levels

Credits “Building Powerful Web Applications in the AWS Cloud” by Louis Columbus
http://softwarestrategiesblog.com/2011/03/10/building-powerful-web-applications-in-the-aws-cloud/

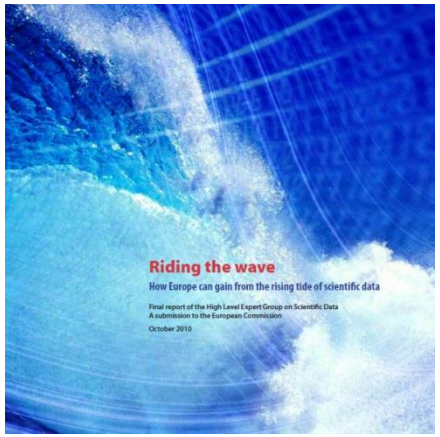
Технологии Больших Даных

Visionaries and Drivers: Seminal works and High level reports



The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, et al.
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

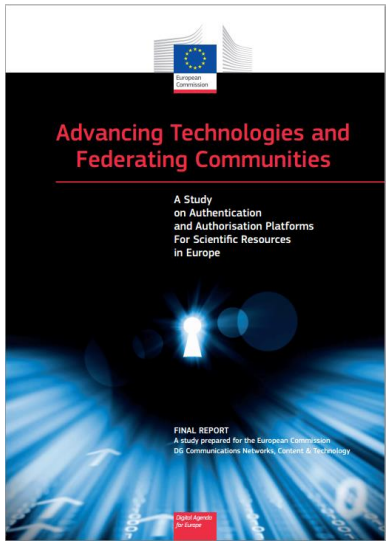


Riding the wave: How Europe can gain from the rising tide of scientific data.

Final report of the High Level Expert Group on Scientific Data. October 2010.
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

RDA
RESEARCH DATA ALLIANCE
Research Data Sharing without barriers
https://www.rd-alliance.org/

NIST Big Data Working Group (NBD-WG)
https://www.rd-alliance.org/



AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe,

TERENA, UvA, LIBER, UinvDeb.

The Fourth Paradigm of Scientific Research

1. Theory and logical reasoning
2. Observation or Experiment
   - E.g. Newton observed apples falling to design his theory of mechanics
   - But Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove theory or model
4. Data-driven Scientific Discovery (aka Data Science)
   - More data beat hypnotized theory

Big Data and Data Intensive Science - The next technology focus

Scientific and Research Data – e-Science

- *Big Data is/has becoming the next buzz word*
  - Not much academic research and papers – Read seminal works, *Dive into blogs and tweets*
- Based on e-Science concept and entire information and artifacts digitising
  - Requires also *new information and semantic models* for information structuring and presentation
  - Requires new research methods using large data sets and data mining
    - Methods to evolve and results to be improved
- Changes the way how the modern research is done (in e-Science)
  - Secondary research, data re-focusing, linking data and publications
- Big Data require ***a new infrastructure*** to support both distributed data (collection, storage, processing) and metadata/discovery services
  - High performance network and computing, distributed storage and access
  - Cloud Computing as native platform for distributed dynamic virtualised (data supporting) infrastructure
  - Demand for trusted/trustworthy infrastructure

Big Data Definitions Overview

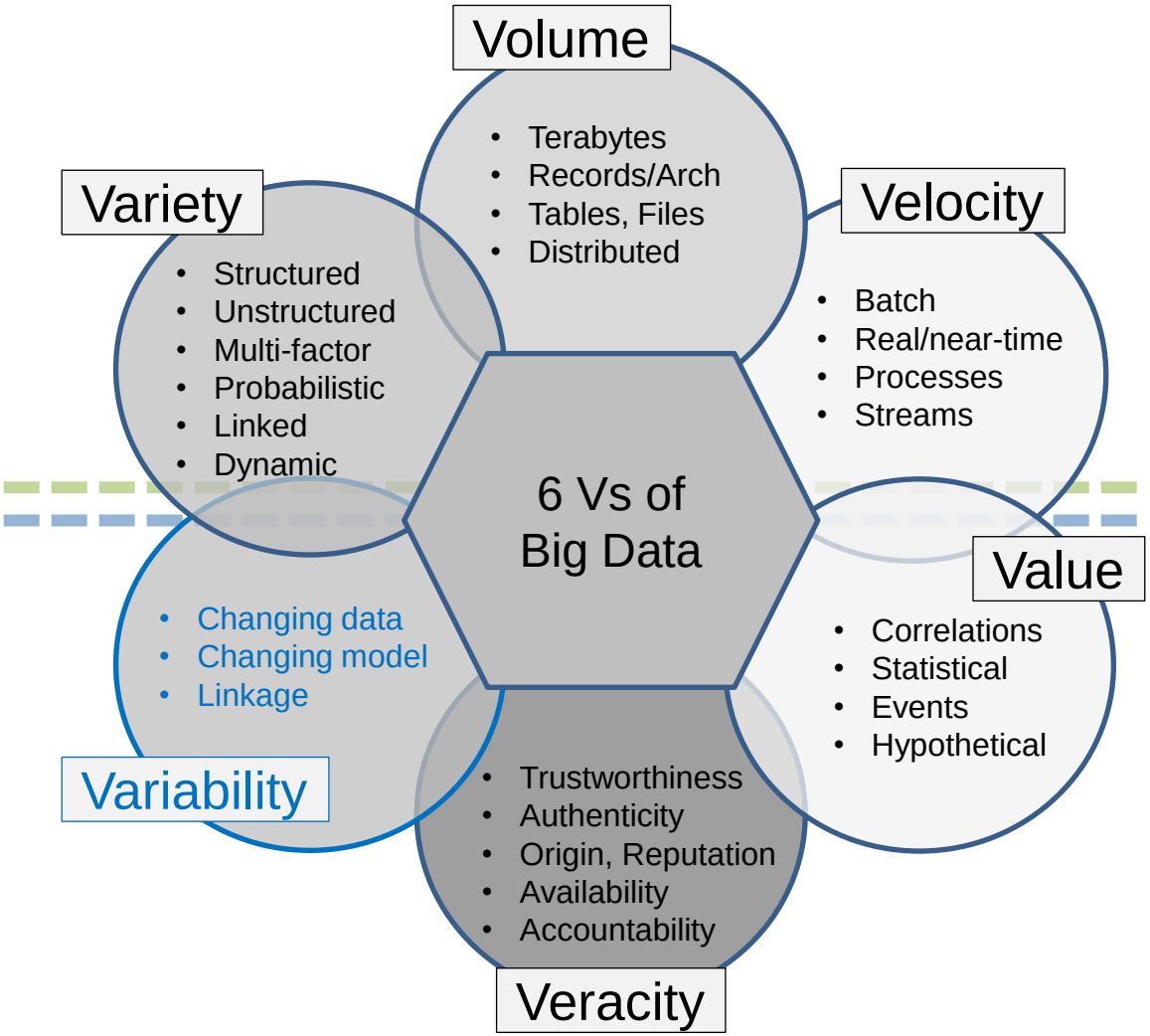
- IDC definition of Big Data (conservative and strict approach) :
  "A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"

- Gartner definition
  Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. http://www.gartner.com/it-glossary/big-data/
  - Termed as 3 parts definition, not 3V definition
- Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.
  - From "The Big Data Long Tail" blog post by Jason Bloomberg (Jan 17, 2013). http://www.devx.com/blog/the-big-data-long-tail.html
- "Data that exceeds the processing capacity of conventional database systems. *The data is too big, moves too fast, or doesn't fit the structures of your database architectures*. To gain value from this data, you must choose an alternative way to process it."
  - Ed Dumbill, program chair for the O'Reilly Strata Conference

- Termed as the Fourth Paradigm *)
  "The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration." (Jim Gray, computer scientist)

**) The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft, 2009.*

Improved: 5+1 V's of Big Data



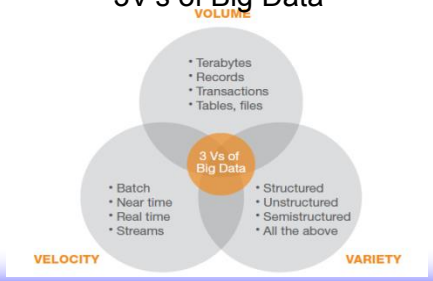
Generic Big Data Properties

- Volume
- Variety
- Velocity

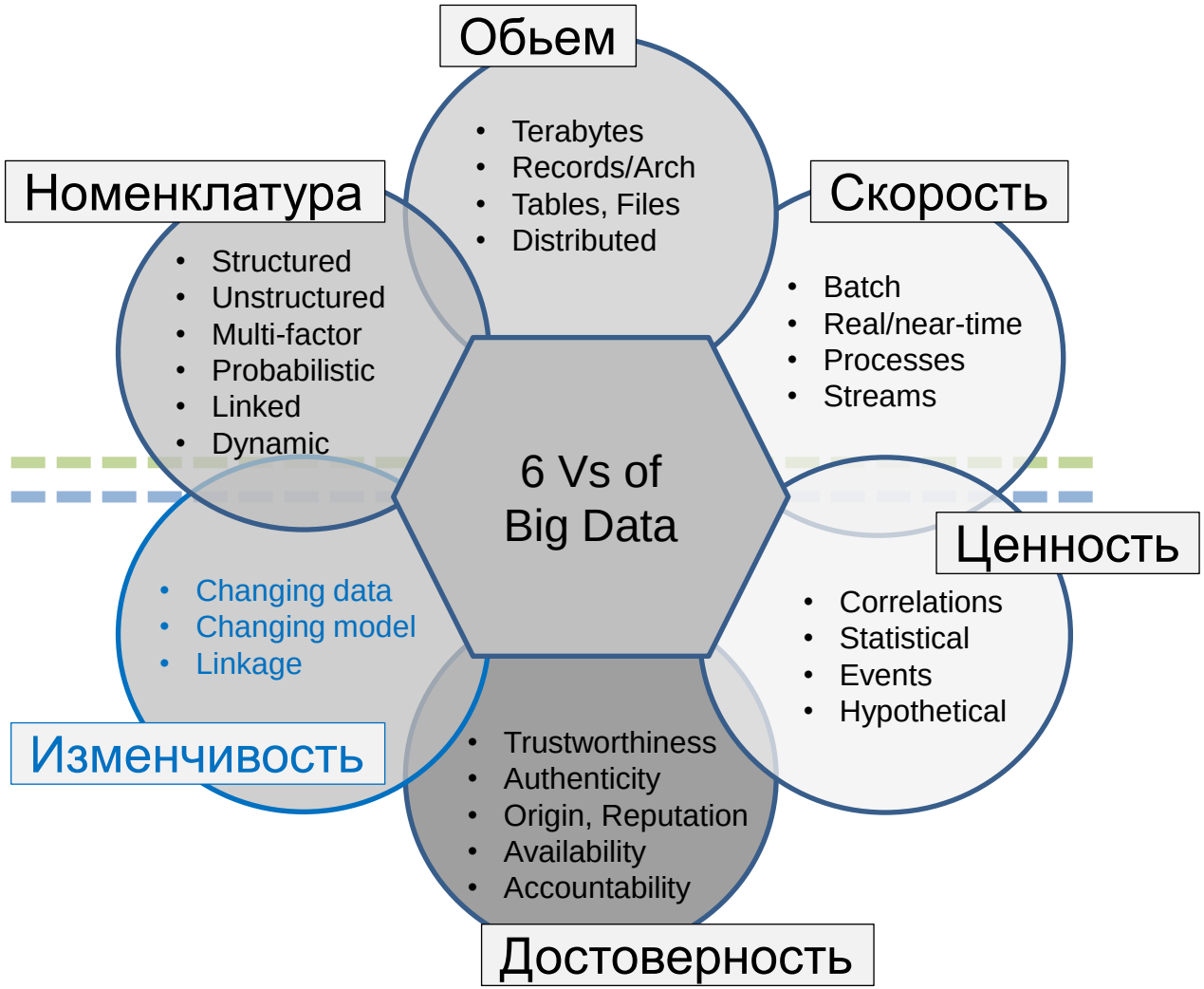
Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data



Improved: 5+1 V's of Big Data



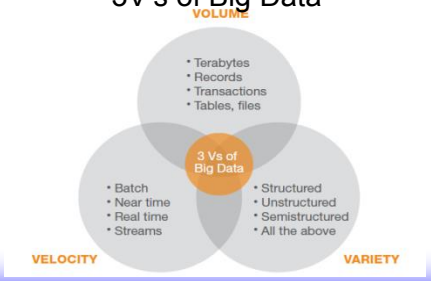
Базовые свойства

- Обьем
- Скорость
- Номенклатура (разнообразие)

Приданные свойства (после обработки)

- Ценность
- Достоверность
- Изменчивость

Commonly accepted 3V's of Big Data



Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
- Volume, Variety, Velocity, Value, Veracity
- Additionally: Data Dynamicity (Variability)

(2) New Data Models
- Data Lifecycle and Variability
- Data linking, provenance and referral integrity

(3) New Analytics
- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
- High performance Computing, Storage, Network
- Heterogeneous multi-provider services integration
- New Data Centric (multi-stakeholder) service models
- New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
- High velocity/speed data capture from variety of sensors and data sources
- Data delivery to different visualisation and actionable systems and consumers
- Full digitised input and output, (ubiquitous) sensor networks, full digital control

Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
- Volume, Variety, Velocity, Value, Veracity
- Additionally: Data Dynamicity (Variability)

(2) New Data Models
- Data linking, provenance and referral integrity
- Data Lifecycle and Variability/Evolution

(3) New Analytics
- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
- High performance Computing, Storage, Network
- Heterogeneous multi-provider services integration
- New Data Centric (multi-stakeholder) service models
- New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
- High velocity/speed data capture from variety of sensors and data sources
- Data delivery to different visualisation and actionable systems and consumers
- Full digitised input and output, (ubiquitous) sensor networks, full digital control

Big Data Definition: From 5V to 5 Parts (2)

Refining Gartner definition

"Big data is (1) high-volume, high-velocity and high-variety information assets that demand (3) cost-effective, innovative forms of information processing for (5) enhanced insight and decision making"

- Big Data (Data Intensive) Technologies are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand (3) cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) (2) new data models (supporting all data states and stages during the whole data lifecycle) and (4) new infrastructure services and tools that allows also obtaining (and processing data) from (5) a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

(1) Big Data Properties: 5V
(2) New Data Models
(3) New Analytics
(4) New Infrastructure and Tools
(5) Source and Target

Big Data Nature: Origin and consumers (target)

Big Data Origin

- Science
- Internet, Web
- Industry
- Business
- Living Environment, Cities
- Social media and networks
- Healthcare
- Telecom/Infrastructure

Data Transformation

Big Data Target Use

- Scientific discovery
- New technologies
- Manufacturing, processes, transport
- Personal services, campaigns
- Living environment support
- Healthcare support
- Social Networking

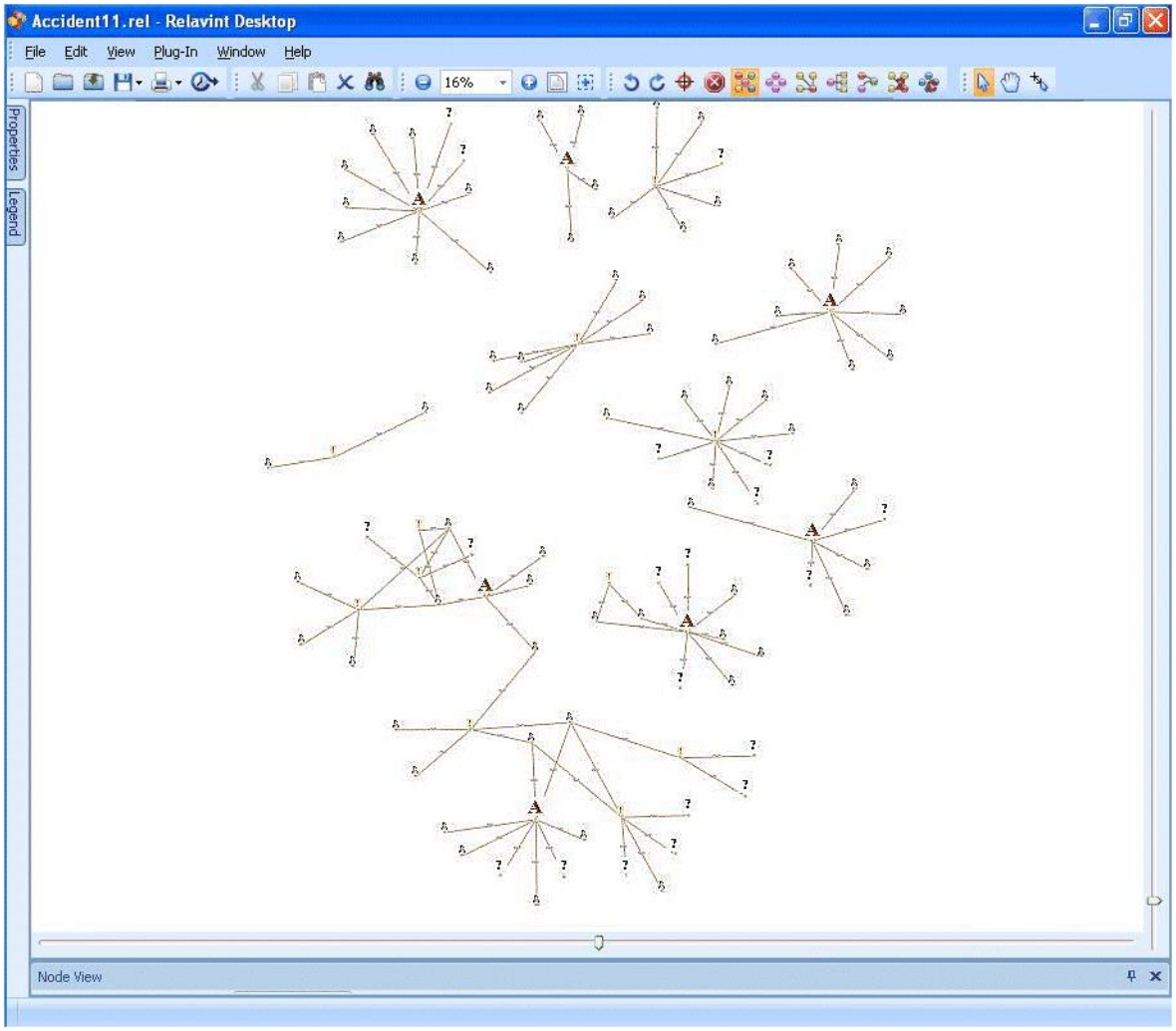
Volume, Velocity, Variety – Examples e-Science

- Volume – Terabyte records, transactions, tables, files.
  - LHC – 5 PB a month (now is under re-construction)
  - LOFAR, SKA – 5 PB every hour, requires processing asap to discard non-informative data
  - Large Synoptic Survey Telescope (LSST) - 10 Petabytes per year
  - Genomic research – x10 TB per individual
  - Earth, climate and weather data
- Velocity – batch, near-time, real-time, streams.
  - LHC ATLAS detector generates about 1 Petabyte raw data per second, during the collision time about 1 ms
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - Biodiversity, Biological and medical, facial research
  - Human, psychology and behavior research
  - History, archeology and artifacts

Volume, Velocity, Variety – Examples Industry

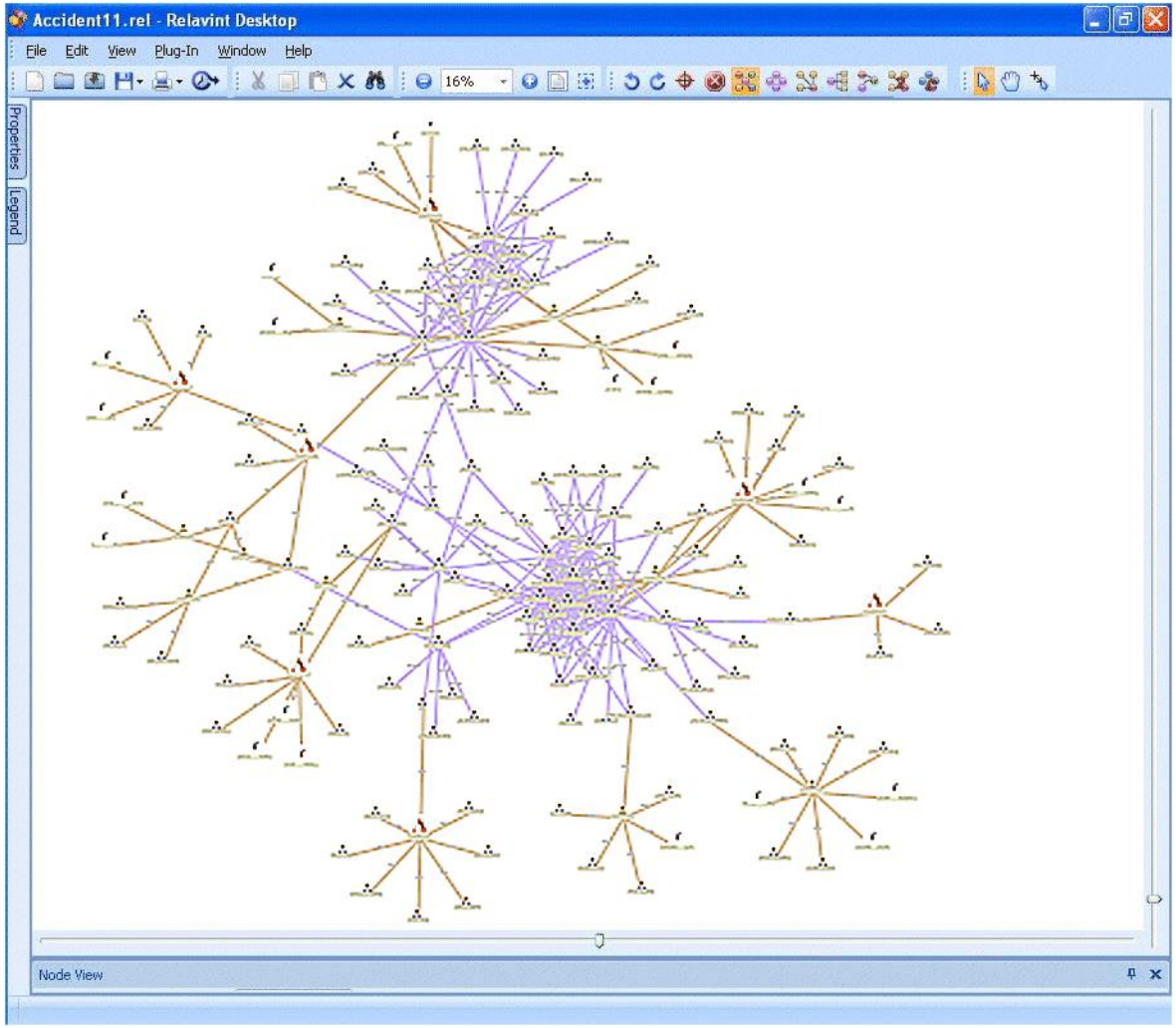
- Volume – Terabyte records, transactions, tables, files.
  - A Boeing Jet engine produce out 10TB of operational data for every 30 minutes they run
  - Hence a 4-engine Jumbo jet can create 640TB on one Atlantic crossing. Multiply that to 25,000 flights flown each day
- Velocity – batch, near-time, real-time, streams.
  - Today's on-line ads serving requires *40ms to respond with a decision*
  - Financial services (i.e., stock quotes feed) need near 1ms to calculate customer scoring probabilities
  - Stream data, such as movies, need to travel at high speed for proper rendering
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - WalMart processes 1M customer transactions per hour and feeds information to a database estimated at 2.5PB (petabytes)
  - There are old and new data sources like RFID, sensors, mobile payments, in-vehicle tracking, etc.

Real Life Graph Analytics (LexisNexis)



The Insurance company data **only finds a connection between two of the seven claims**, and only identified one other claim as being weakly connected.

Real Life Graph Analytics (LexisNexis)



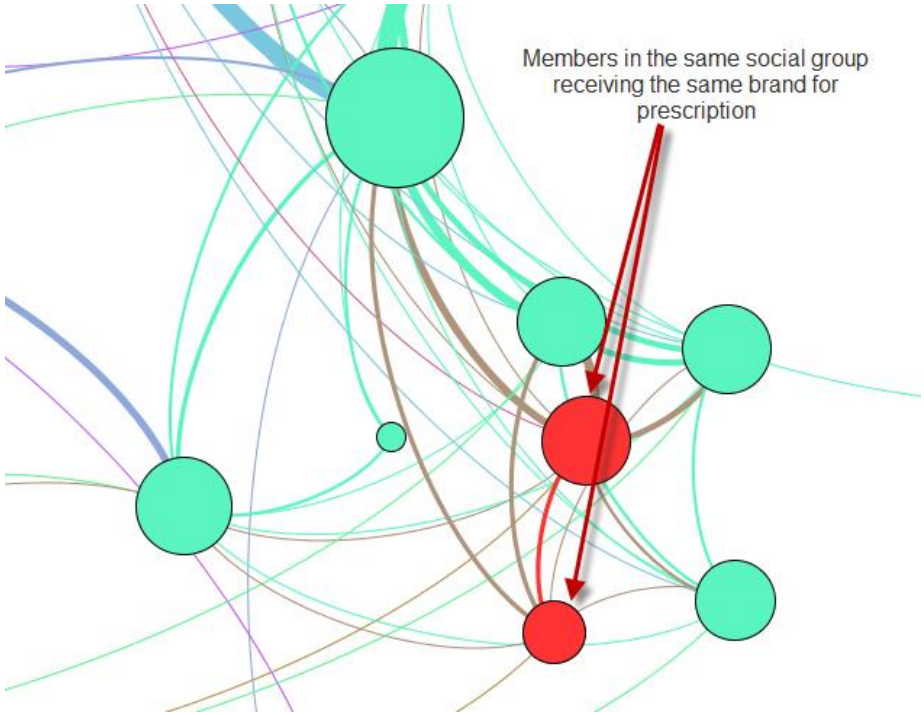
Task

- After adding the LexID to the carrier Data, LexisNexis HPCC technology then explored 2 additional degrees of relative separation

Result

- The results showed **two family groups interconnected on all of these seven claims**.
- The links were much stronger than the carrier data previously supported.

Social Graph and Prescription (LexisNexis)



Task

- Given a large set of prescriptions. Calculate normal social distributions of each brand and detect where there is an unusual socialization of prescriptions and services.

Result

- The analysis detected social groups that are sourcing Vicodin and other schedule drugs. Identifies prescribers and pharmacies involved to help the insurer focus investigations and intervene strategically to mitigate risk.

Big Data technology drivers (1)

- Modern e-Science in search for new knowledge
  - Scientific experiments and tools are becoming bigger and heavily based on data processing and mining
- Traditional data intensive industry
  - Genomic research, drugs development, Healthcare
  - High-tech industry, CAD/CAM, weather/climate, etc.
- Intelligence and security
- Network/infrastructure management
  - Network monitoring, Intrusion detection, troubleshooting

Big Data technology drivers (2)

- Consumer facing companies like Google and Facebook have driven many of the recent advances in Big Data efficiency
  - Facebook has some 900+ million users and is still growing
  - Google handles number of search queries at 3 billion per day
  - Twitter handles some 400 million tweets per day count for 12 terabytes per day
    - Used also for market sentiments prediction
  - Power companies: process up to 350 billion annual meter readings to better predict power consumption
- Processes/activity data recording and analysis
  - Flight data, log data, intelligence, traffic
- **Business (retail) uses Big Data technologies "to search" for customers**
  - Modern business concept (multi-channel) of delivering directly to customers requires prediction of customer behavior
    - Data volumes – What cause(s) and what effect?
  - Big Data gives companies a fighting chance in the battle over the customer

Big Data technology drivers (3) - Advertising

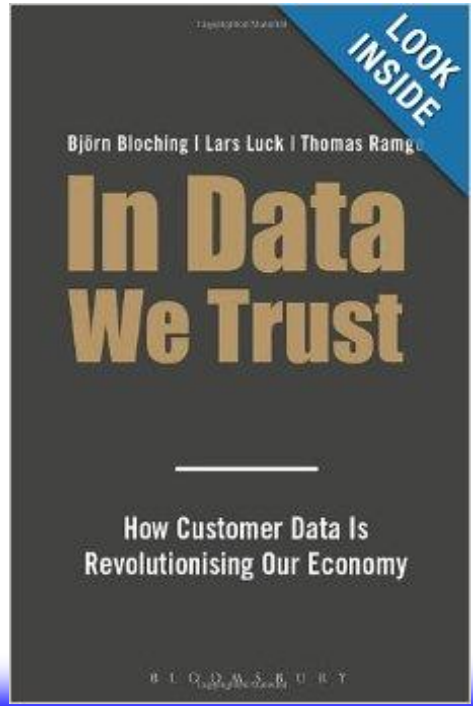
- "… this new course of big data, gleaned from a wealth of unstructured information on the web, has the ability to turn advertising on its head— at least enough to make media people rethink algorithms for maximizing performance." *HessieJ.com*
  - *Traditional Ad Model: User profiles*
  - *More Sophisticated Ad Model: Behavioral targeting - "smart ads"*
  - *Future Ad Model: Enter Social Data*
- Example:
  - Mary Brown searches for information about a future trip to Mallorca
    - She also goes to travel sites, reads hotel reviews and has excitedly spoken to close friends on Twitter and Facebook about her plans and preparations
  - Now we have not only recent behavioral activity *where she's been on the internet*, but we also are aware of her conversations that validate her behavior
  - It is safe to assume that Mary will "definitely" be going to Mallorca
  - What this information does for a travel company?
    - They now have ***MORE*** information on that user that will allow them to not only *serve* an ad, or even *respond* to that user with relevant offers, but ***DO*** so with a certain degree of confidence that Mary will at the very least click on the ad.

Big Data technology drivers (3a) – Service Delivery

- Consumer products and services delivery
  - Netflix already captures movie genre preferences by the user and makes recommendations based on recent shows/movies watched
    - Announced $2mln prize for effective customer targeting in 2003
    - Netflix recommender system in use as a reference technology implementation
  - It is already capturing which devices the user is watching recent programs/shows and when
    - Marrying that data with GetGlue (news feed on movies), for example, validates the original information and supplements the usage information
  - Combined and correlating, allows Netflix to optimize the movie offering to keep you a satisfied customer
  - It can also capture the comments and shares from those watching the movie in order *to drive messaging* to attract new users

Big Data technology drivers (3b) – Managing public campaigns, e.g. election, public relations

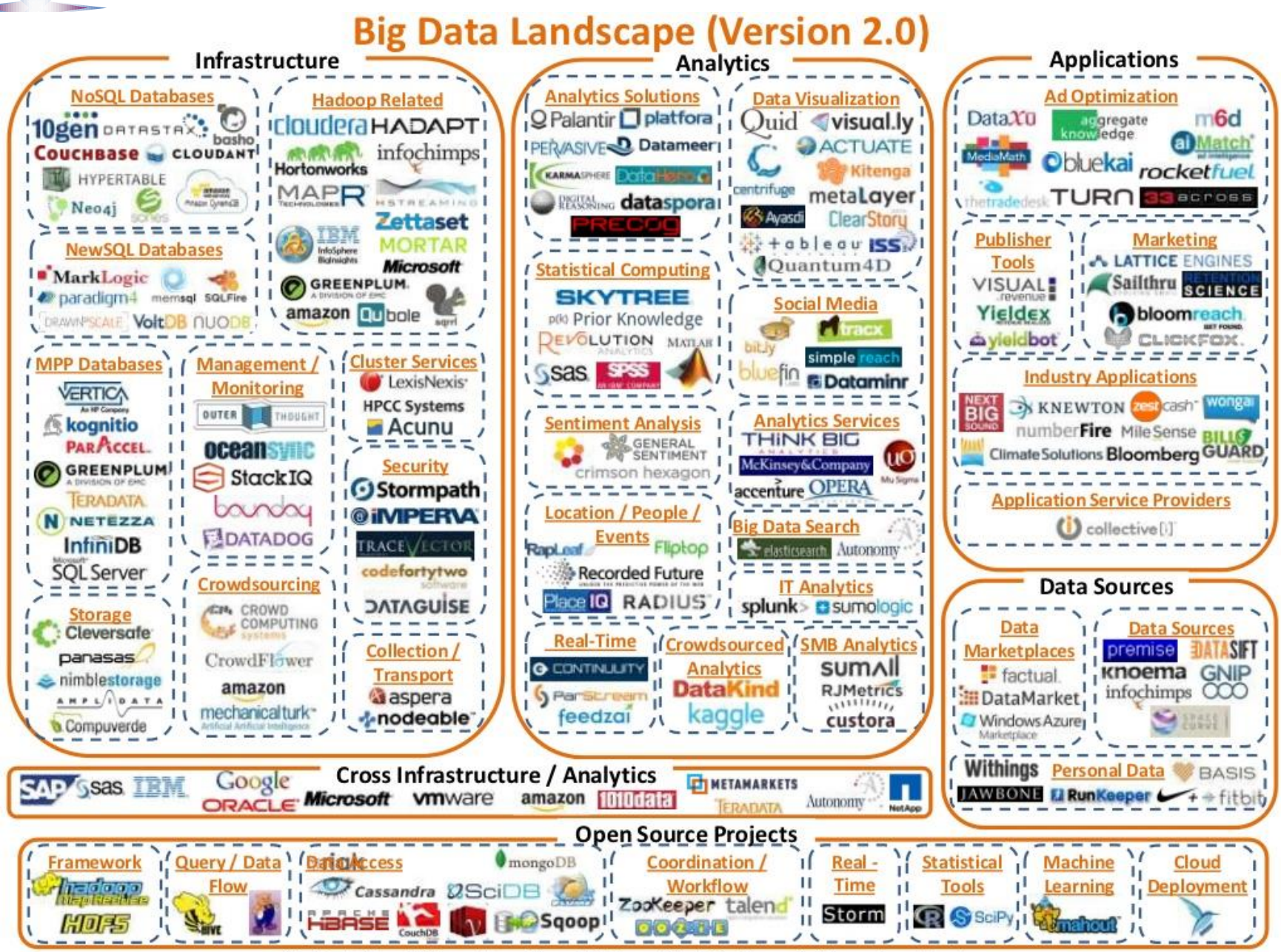
- The rise of public opinion stored in platforms like Twitter, Google, Facebook, etc. provide enough intelligence to influence the campaign development, timing, geography and even the colour of the campaign signs
  - Twitter was a major source of data aggregation for the Republican Race in the US
  - Multimillion-dollar contract for data management and collection services awarded May 1, 2013 to Liberty Work to build advanced list of voters
    - Article “In Data we trust” by T.Edsall in The New York Times
  - Book: In Data We Trust: How Customer Data is Revolutionising Our Economy (Aug 2012)
    - A strategy for tomorrow's data world



Big Data technology drivers (4) - Emerging

- Social media itself – share and socialise/collaborate
  - Facebook, Twitter, YouTube, Flickr, etc.
- Workplace improvement
  - Means more data will be collected and monitored on the personnel
- Healthcare, health/physiological and medical information
  - Human health monitoring – not just for ill or aged people
    - Early diagnostics, proactive care advising

http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/



Foreseen Big Data Innovations in 2013+

- Cloud-Based Big Data Solutions
  - Amazon's Elastic Map Reduce (EMR) is a market leader
  - Expected new innovative Big Data and Cloud solutions
- Real-Time Hadoop
  - Google's Dremel-like solutions that will allow real-time queries on Big Data and be open source
- Distributed and deep Machine Learning
  - Mahout iterative scalable distributed back propagation machine learning and data mining algorithm
  - New algorithms Jubatus, HogWild
- Big Data Appliances (also for home)
  - Raspberry Pi and home-made GPU clusters
  - Hardware vendors (Dell, HP, etc.) pack mobile ARM processors into server boxes
  - Adepteva's Parallella will put a 16-core supercomputer into range of $99
- Easier Big Data Tools
  - Open Source and easy to use drag-and-drop tools for Big Data Analytics to facilitate the BD adoption
  - Commercial examples: Radoop = RapidMiner + Mahout, Tableau, Datameer, etc.
  - LexisNexis: from ECL (Enterprise Control Language) to KEL (Knowledge Engineering Language)

Source: Big Data in 2013 by Mike Guattieri, Forrester

NIST Big Data Working Group (NBD-WG)

- Deliverables target – September 2013
  - 26 September – initial draft documents
  - 30 September – Workshop and F2F meeting
- Activities: Conference calls every day 17-19:00 (CET) by subgroup - http://bigdatawg.nist.gov/home.php
  - Big Data Definition and Taxonomies
  - Requirements (chair: Geoffrey Fox, Indiana Univ)
  - Big Data Security
  - Reference Architecture
  - Technology Roadmap
- BigdataWG mailing list and useful documents
  - Input documents http://bigdatawg.nist.gov/show_InputDoc2.php
  - Big Data Reference Architecture http://bigdatawg.nist.gov/_uploadfiles/M0226_v7_2611176301.docx
  - Big Data Architectures Survey http://bigdatawg.nist.gov/_uploadfiles/M0151_v2_7447424902.docx
  - Requirements based on 51 usecases http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx

Defining Big Data Architecture Framework

- Existing attempts don't converge to something consistent: ODCA, TMF, NIST
  - See http://bigdatawg.nist.gov/_uploadfiles/M0151_v2_7447424902.docx
- Architecture vs Ecosystem
  - Big Data undergo a number of transformations during their lifecycle
  - Big Data fuel the whole transformation chain
    - Data sources and data consumers, target data usage
  - Multi-dimensional relations between
    - Data models and data driven processes
    - Infrastructure components and data centric services
- Architecture vs Architecture Framework
  - Separates concerns and factors
    - Control and Management functions, orthogonal factors
  - Architecture Framework components are inter-related

Big Data Architecture Framework (BDAF) (1)

(1) Data Models, Structures, Types
- Data formats, non/relational, file systems, etc.

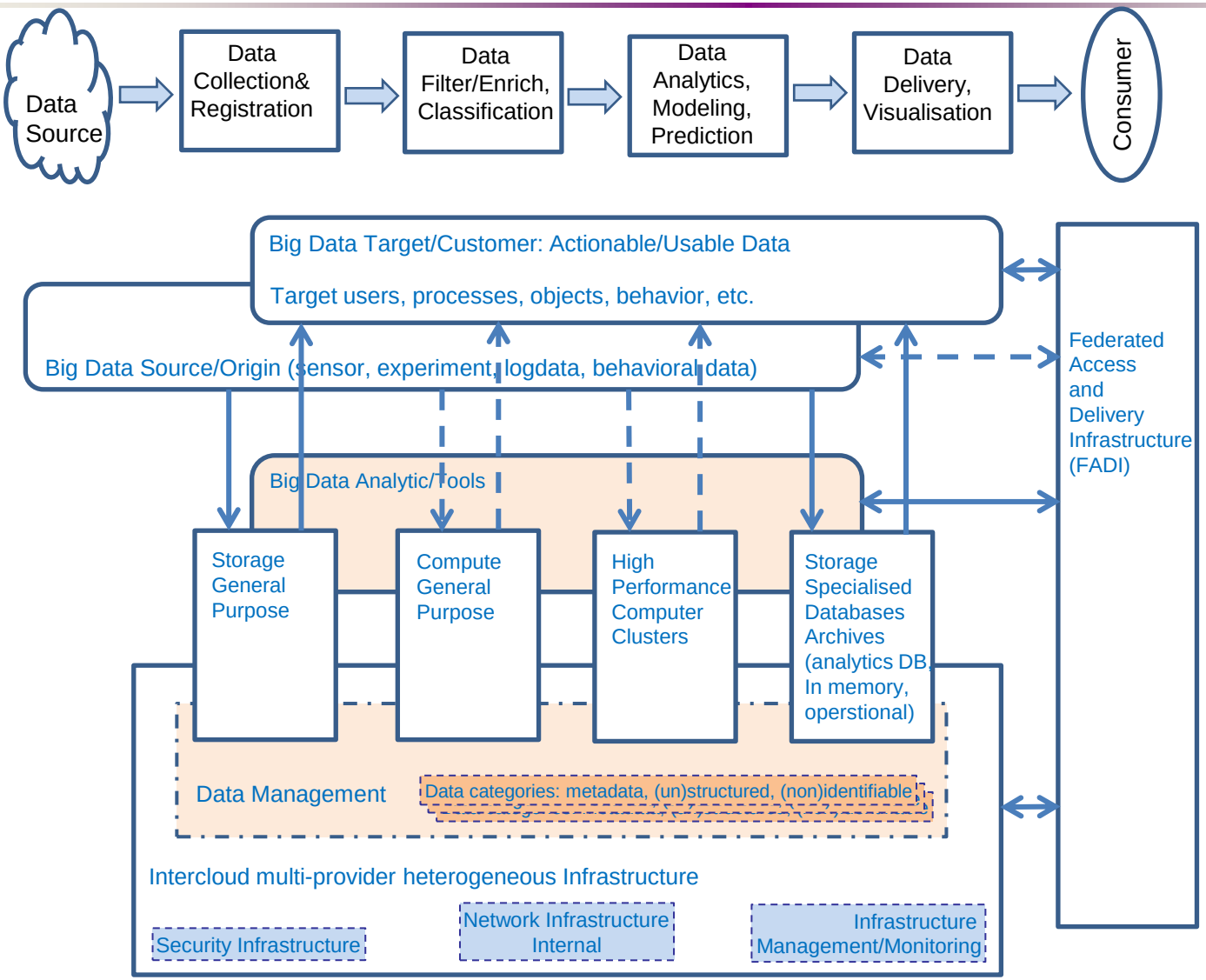
(2) Big Data Management
- Big Data Lifecycle (Management) Model
  - Big Data transformation/staging
- Provenance, Curation, Archiving

(3) Big Data Analytics and Tools
- Big Data Applications
  - Target use, presentation, visualisation

(4) Big Data Infrastructure (BDI)
- Storage, Compute, (High Performance Computing,) Network
- Sensor network, target/actionable devices
- Big Data Operational support

(5) Big Data Security
- Data security in-rest, in-move, trusted processing environments

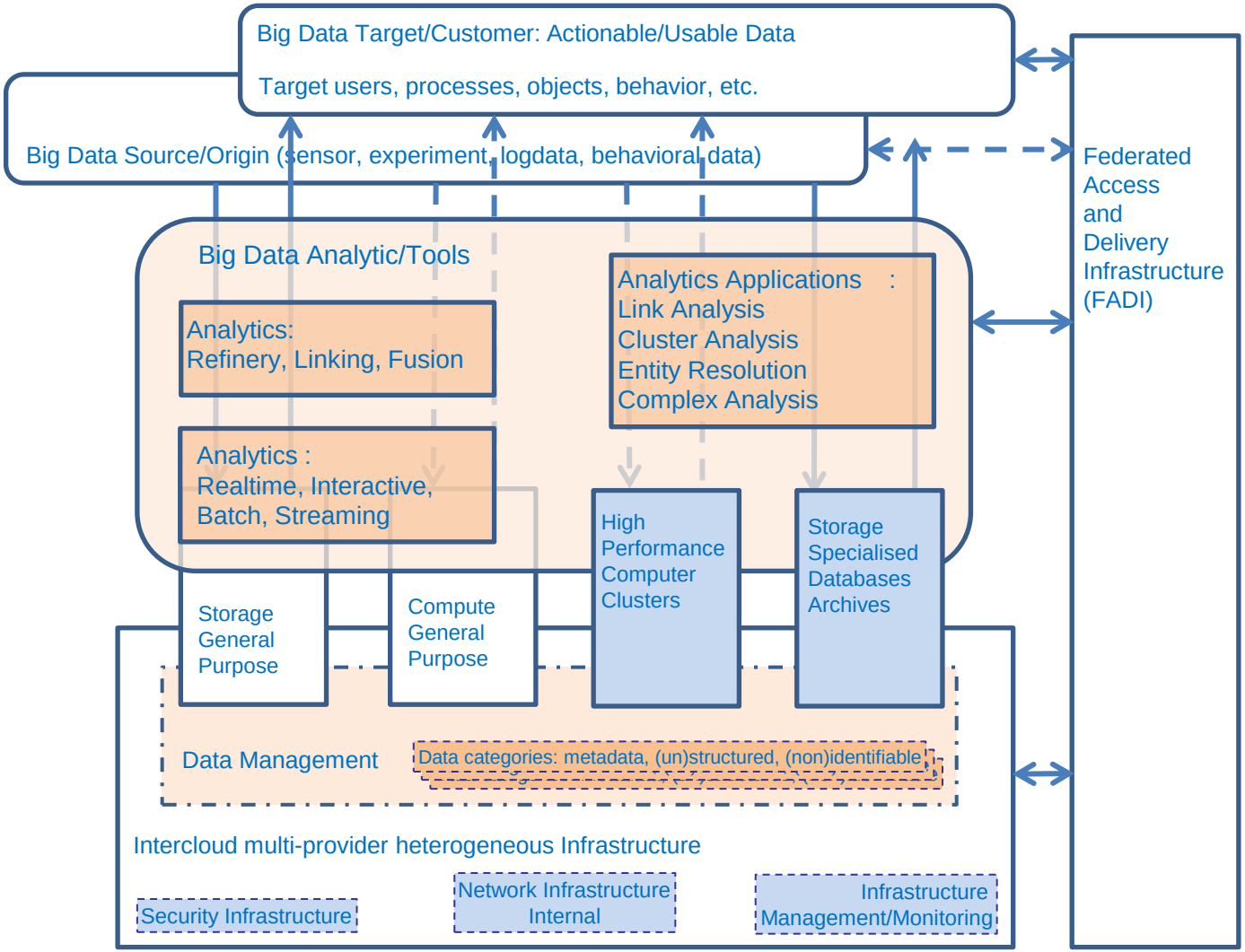
Big Data Ecosystem: Data, Transformation, Infrastructure



General BDI services and components

- Data management infrastructure and tools
- Registries, search/indexing, ontologies, schemas, namespace
- Collaborative Environment (user/groups managements)
- Heterogeneous multi-provider Inter-cloud infrastructure
  - Compute, Storage, Network (provisioned on-demand dynamically scaling)
  - Federated Access and Delivery Infrastructure (FADI)
- Advanced high performance (programmable) network
- Security infrastructure (access control, Identity and policy management, confidentiality, privacy, trust)

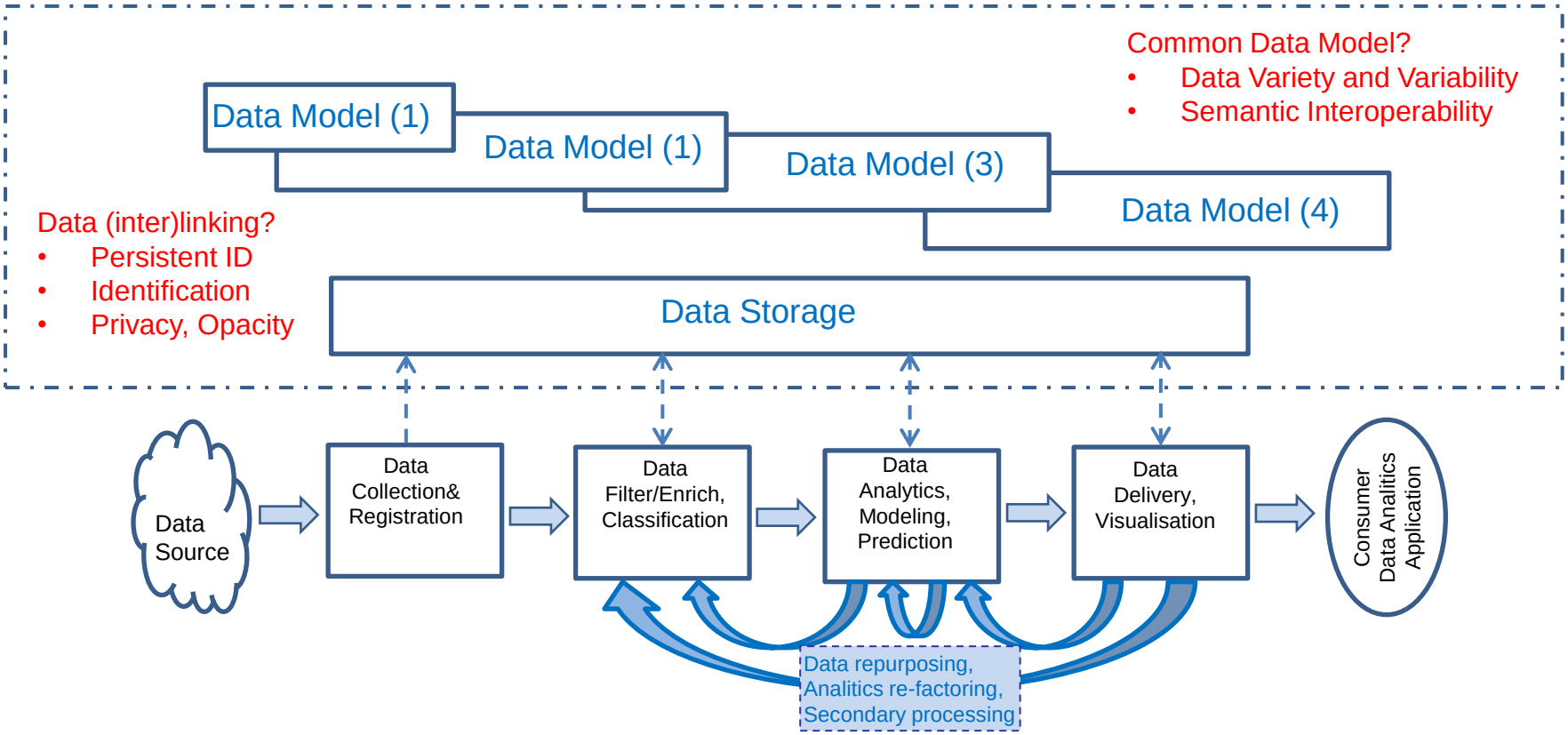
Big Data Infrastructure and Analytic Tools



Big Data Analytics Infrastructure

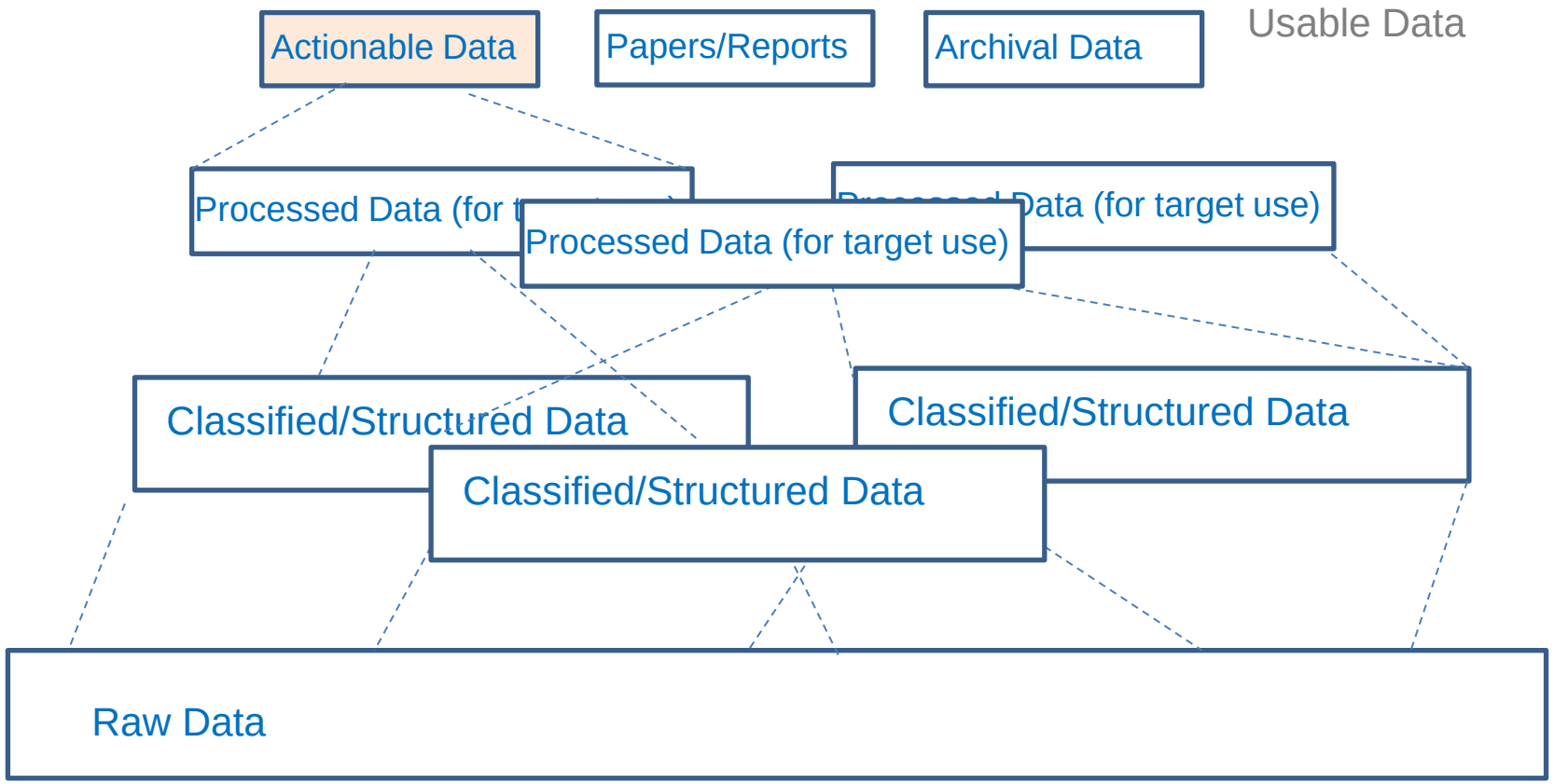
- High Performance Computer Clusters (HPCC)
- Specialised Storage, Distributed/Replicated, Archives, Databases, SQL/NoSQL
- Big Data Analytics Tools/Applications
- Analytics/processing: Real-time, Interactive, Batch, Streaming
- Link Analysis, Graph analysis
- Cluster Analysis
- Entity Resolution
- Complex Analysis

Data Transformation/Lifecycle Model



- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifier
  - Traceability vs Opacity
  - Referral integrity

Evolutional/Hierarchical Data Model



- Common Data Model?
- Data interlinking?
- Fits to Graph data type?
- Metadata

- Referrals
- Control information
- Policy
- Data patterns

Data Science: Research and Education

Horizon2020: Challenge (2.1): Development, deployment and operation of e-infrastructures

Call on Specific Challenge 2.1: Development, deployment and operation of e-infrastructures (based on discussion draft summer 2013)

CHALLENGE 1 – High Performance Computing (HPC)

CHALLENGE 2 - CONNECTIVITY

- **Topic 4: Research and Education Networking – GÉANT**

CHALLENGE 3 - DATA

- Topic 5: Community data services
- Topic 6: Managing, preserving and computing with big research data
- Topic 7: e-Infrastructure for Open Access
- Topic 8: Towards global data e-infrastructures - RDA

CHALLENGE 4 – e-INFRASTRUCTURE INTEGRATION

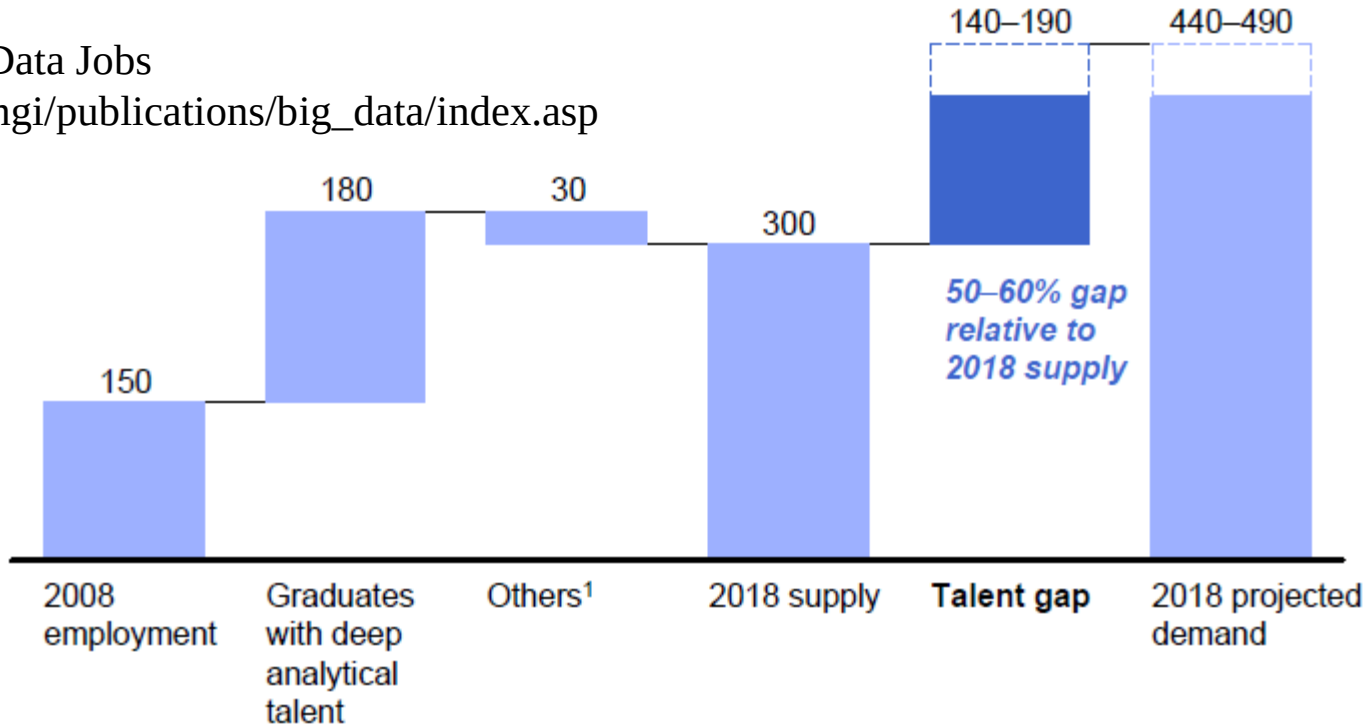
- Topic 9: e-Infrastructures for virtual research environments (VRE)
- Topic 10: Provisioning of core services across e-Infrastructures
- **Topic 11: Skills and new professions for e-infrastructures**

CHALLENGE 5 – POLICY AND INTERNATIONAL

- Topic 12: Policy development and international cooperation

Data Scientist: New Profession and Opportunities

McKinsey Institute on Big Data Jobs
http://www.mckinsey.com/mgi/publications/big_data/index.asp

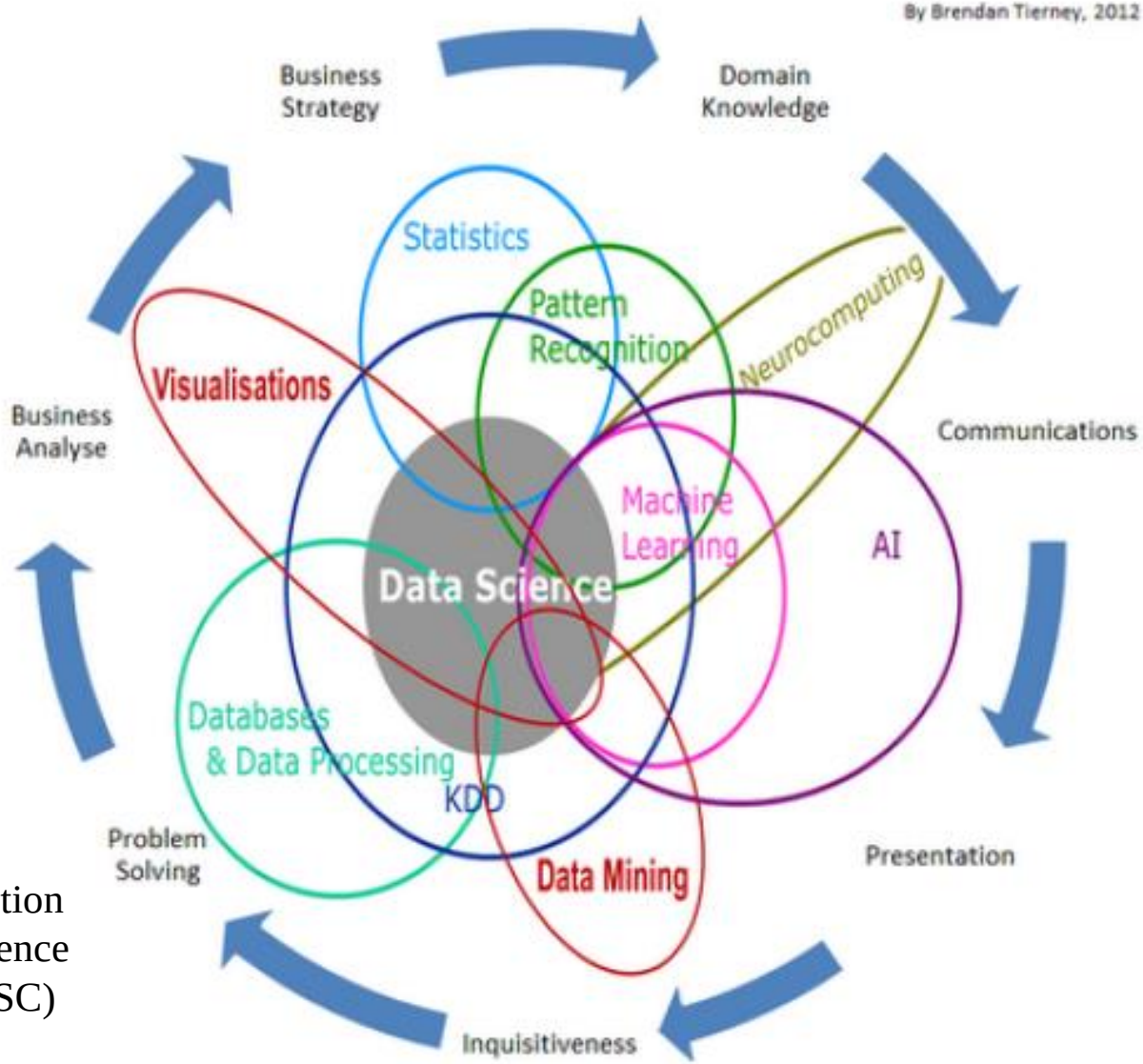


- There will be a shortage of talent necessary for organizations to take advantage of Big Data.
  - By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as
  - 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions

SOURCE:US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey analysis

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Slide from the presentation Demystifying Data Science (by Natasha Balac, SDSC)

Data Science: Data driven processes

- Store and process
  - Large scale databases
  - Software Engineering
  - System/network engineering
- Analyse and model
  - Reasoning
  - Knowledge representation
  - Multimedia Retrieval
  - Modelling and simulation
  - Machine Learning
  - Information Retrieval
- Understand and design
  - Decision theory
  - Visual analytics
  - Perception Cognition

Credits: DSRC, UvA http://sbt.science.uva.nl/wordpress/?page_id=109

Strata Survey Skills and Data Scientist Self-ID

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Analysing the Analysers. O'Reilly Strata Survey – Harris, Murphy&Vaisman, 2013

- Based on how data scientists think about themselves and their work
- Identified four Data Scientist clusters

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Skills and Self-ID Top Factors

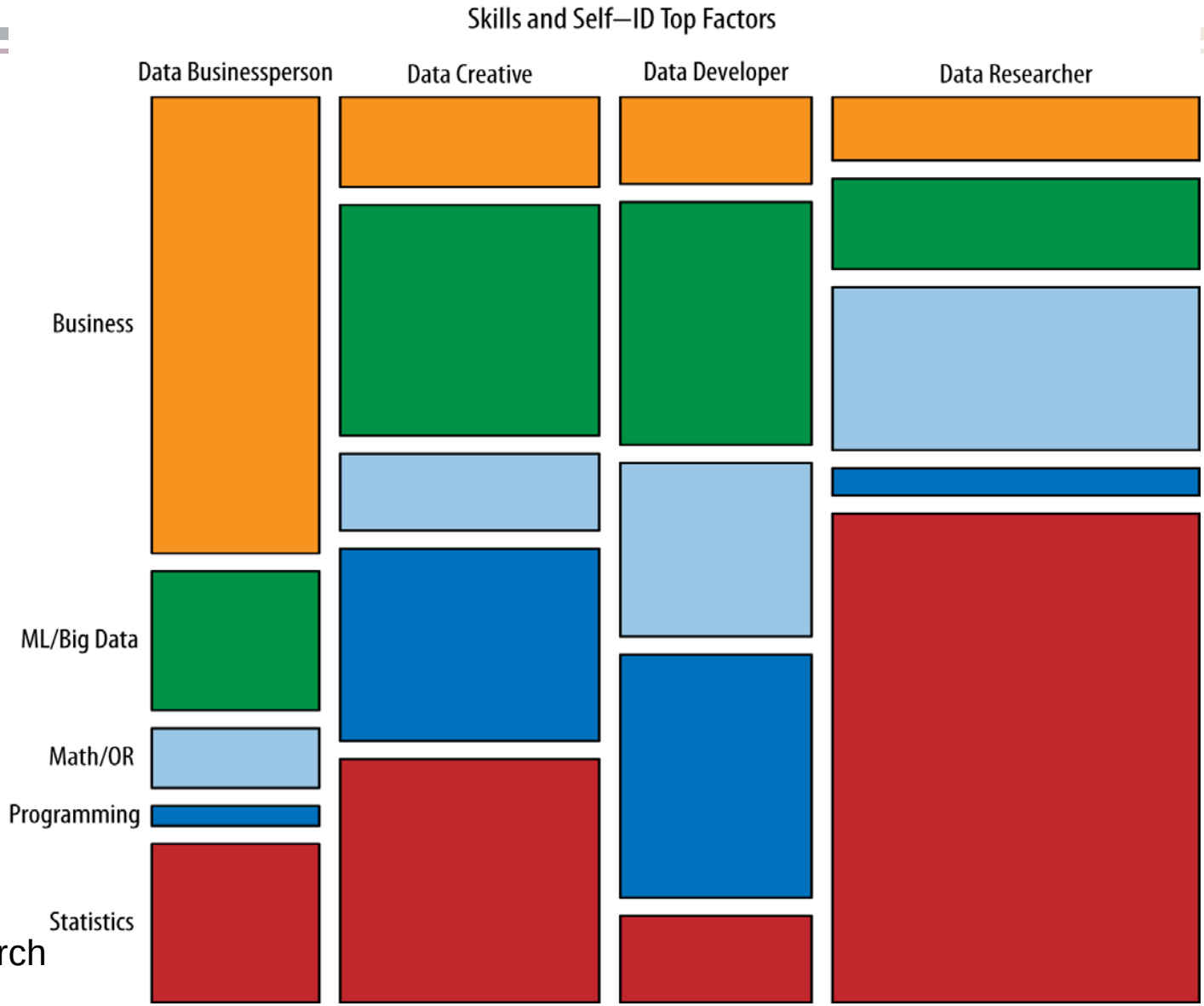
Business

ML/BigData

Math/OR

Programming

Statistics



ML – Machine Learning

OR – Operations Research

Key to a Great Data Scientist

Technical skills (Coding, Statistics, Math)

+ Commitment + Creativity

+ Intuition

+ Presentation Skills

+ Business Savvy

= Great Data Scientist!

- How Long Does It Take For a Beginner to Become a Good Data Scientist?
  - *3-5 years according to* KDnuggets survey [278 votes total]

Summary Big Data

- *Researching, learning mastering Big Data domain is a Big Data problem itself*
- Cloud Computing as a native platform for Big Data
  - Acceptance of clouds will grow, so demand for specialists
- Demand for advanced high performance network will remain and grow
- New generically data centric models are required
- New distributed data processing and analytics computing models to be developed/re-factored
- Data Scientist is a new focus for talents search by companies

Теми презентації та питання для дискусії

1. Характеристики і можливості хмарних технологій, тенденції розвитку та стандартизація.
2. Приклади використання та типи впровадження комп'ютерних хмар: корпоративні, публічні, комунальні; міграція корпоративної IT інфраструктури на хмарну платформу, необхідні передумови і рівень "зрілості", переваги віртуалізації сервісів і ресурсів.
3. Законодавча та регуляторна база в Європі, програми підтримки впровадження хмар в Європі.
4. Глобальні провайдери хмарних послуг і ресурсів: Amazon AWS, Microsoft Azure, GoogleCloud: можливості, послуги, засоби розробки.
5. Великі Дані: Об'єм, Швидкість, Номенклатура, Мінливість, Цінність , Достовірність (Volume, Velocity, Variety, Variability, Value, Veracity).
6. Великі Дані та бізнес- аналітика: приклади використання і нові можливості.
7. Проблеми Великих Даних: зберігання, передача, обробка, контроль доступу, захист даних і персональної інформації.
8. Нові спеціальності для Хмарних технологій та Великих Даних: підготовка фахівців, тренінг та освіта.

Questions and Discussion

Useful links

- SNE Technical Report on Architecture Framework and Components for the Big Data Ecosystem. Draft Version 0.2, 12 September 2013
  http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf
- Addressing Big Data Issues in Scientific Data Infrastructure. By Yuri Demchenko, et al. Paper at IEEE CTS 2013 Conf., May 20-24, 2013, San Diego, USA
  http://www.uazone.org/demch/papers/bddac2013-bigdata-infrastructure-v06.pdf
- BoF on Big Data Challenges for NREN's organised at TNC2013
  https://tnc2013.terena.org/core/event/15
- NIST Big Data Working Group (NBD-WG) http://bigdatawg.nist.gov/home.php
  - Input documents http://bigdatawg.nist.gov/show_InputDoc2.php
  - Big Data Reference Architecture
    http://bigdatawg.nist.gov/_uploadfiles/M0226_v7_2611176301.docx
  - Big Data Architectures Survey
    http://bigdatawg.nist.gov/_uploadfiles/M0151_v2_7447424902.docx
  - Requirements for 51 usecases
    http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx

NIST Documents on Cloud Computing (1)

- [NIST CC] NIST SP 800-145, “A NIST definition of cloud computing”, [online] Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
- [NIST CCRA] NIST SP 500-292, Cloud Computing Reference Architecture, v1.0. [Online] http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505
- [NIST Synopsis] DRAFT NIST SP 800-146, Cloud Computing Synopsis and Recommendations. [online] Available: http://csrc.nist.gov/publications/drafts/800-146/Draft-NIST-SP800-146.pdf
- Draft SP 800-144 Guidelines on Security and Privacy in Public Cloud Computing. [online] Available: http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf
- NIST SP 500-299: NIST Cloud Computing Security Reference Architecture. [online] http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/CloudSecurity/NIST_Security_Reference_Architecture_2013.05.15_v1.0.pdf

NIST Documents on Cloud Computing (2)

- NIST SP 500-293: US Government Cloud Computing Technology Roadmap Volume 1, High-Priority requirements to Further USG Agency Cloud Computing Adoption. http://www.nist.gov/itl/cloud/upload/SP_500_293_volumeI-2.pdf
- NIST SP 500-293: US Government Cloud Computing Technology Roadmap Volume II, Useful Information for Cloud Adopters (Draft) http://www.nist.gov/itl/cloud/upload/SP_500_293_volumeII.pdf
- NIST SP 500-293: US Government Cloud Computing Technology Roadmap Volume III, Technical Considerations for USG Cloud Computing Deployment Decisions (Draft). http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/RoadmapVolumeIIIWorkingDraft/NIST_cloud_roadmap_VIII_draft_110311.pdf