Big Data and Data Intensive (Science) Technologies

Yuri Demchenko
System and Network Engineering, UvA
Outline

- Big Data and Data Intensive Science as a new technology wave
- Big Data 6 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
- Standardisation in Big Data
- Big Data and Cloud
- Big Data: Data Protection and Privacy
  - US-EU Safe Harbor Framework
- Data Scientist: New profession and need for Education & Training
- Summary and take away
Gartner Technology Hypercycle (2013)

Source: [http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp](http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp)
Gartner Technology Hypercycle (2014)

Source: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
years to mainstream adoption

benefit
- transformational
  - less than 2 years
  - 2 to 5 years
  - 5 to 10 years
  - more than 10 years
- high
- moderate

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 entered 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
Visionaries and Drivers: Seminal works and High level reports

The Fourth Paradigm: Data-Intensive Scientific Discovery.

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

AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe, TERENA, UvA, LIBER, UinvDeb.

NIST Big Data Working Group (NBD-WG)
https://www.rd-alliance.org/
The Fourth Paradigm of Scientific Research

1. Theory and logical reasoning
   - Based on observation and *imagination* of ancient philosophers

2. Observation or Experiment
   - E.g. Newton observed apples falling from the tree to design his theory of mechanics, verified with other observations, validated with experiments
   - 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 Intensive Science)
   - e-Science as computing and Information Technologies empowered science
   - More data beat hypnotized theory
     * But not to use “microscope blindly”*
The Fourth Paradigm of Scientific Research

0. Religion and esoteric beliefs
   – Iris unveiled by Helen Blavatsky study of thinking and theories in ancient world

1. Theory and logical reasoning
   – Based on observation and imagination of ancient philosophers

2. Observation or Experiment
   – E.g. Newton observed apples falling from the tree to design his theory of mechanics, verified with other observations, validated with experiments
   – Galileo 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 Intensive Science)
   – e-Science as computing and Information Technologies empowered science
   – More data beat hypnotized theory
     • But not to use “microscope blindly”
Scientific and Research Data – e-Science

• *Big Data is/has becoming a new 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/](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.

- “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)

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
- Volume
- Velocity
- Variety
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

**Big Data Target Use**
- Scientific discovery
- New technologies
- Manufacturing, processes, transport
- Living environment support
- Healthcare support
- Personal services, campaigns, media
- Social Networks
- Intelligence

*Data Transformation*
Volume, Velocity, Variety – Examples Science

• Volume – Terabyte records, transactions, tables, files.
  – LHC (Large Hadron Collider)
    • 5 PB a month (now is under re-construction)
  – LOFAR (Low Frequency Array), SKA (Square Kilometer Array)
    • 5 PB every hour, requires processing asap to discard non-informative data
  – Large Synoptic Survey Telescope (LSST)
    • 10 Petabytes per year of the complex interlinked hierarchical data
  – 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
Collectively “Long Tail” science is generating a lot of data
- Estimated as over 1PB per year and it is growing fast with the new technology proliferation
- Big Data and Data Science technologies development facilitates collecting more data and using Big Data analytics tools

80-20 rule: 20% users generate 80% data but not necessarily 80% knowledge

Source: Dennis Gannon (Microsoft)
NIST Big Data Workshop, 2012
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.
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.
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
Technology Loop and Sustainable Big Data
Technology Drivers

- Technology loop (known as Jevons Paradox)
  - Increased efficiency to process current demand (for data analysis) will create new uses and increase demand even more

Elastic Demand for Work:
A doubling of fuel efficiency more than doubles work demanded, increasing the amount of fuel used. Jevons paradox occurs.

“… 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 -> Sentiments analysis

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
    • Fitness gadgets and Apps
From Big Data to All-Data – Paradigm Change

• Breaking paradigm changing factor
  – Data storage and processing
  – Security
  – Identification and provenance

• Traditional model
  – BIG Storage and BIG computer with FAT pipe
  – Move compute to data vs Move data to compute

• New Paradigm
  – Continuous data production
  – Continuous data processing
  – DataBus as Data container and Protocol
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
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
NIST Big Data Working Group (NBD-WG) and ISO/IEC JTC1 Study Group on Big Data (SGBD)

  - Built on experience of developing the Cloud Computing standards fully accepted by industry
- Set of documents delivered September 2014 (to be published as NIST Special Publication documents) - [http://bigdatawg.nist.gov/V1_output_docs.php](http://bigdatawg.nist.gov/V1_output_docs.php)
  - Volume 1: NIST Big Data Definitions
  - Volume 2: NIST Big Data Taxonomies
  - Volume 3: NIST Big Data Use Case & Requirements
  - Volume 4: NIST Big Data Security and Privacy Requirements
  - Volume 5: NIST Big Data Architectures White Paper Survey
  - Volume 6: NIST Big Data Reference Architecture
  - Volume 7: NIST Big Data Technology Roadmap

- NBD-WG defined 3 main components of the new technology:
  - Big Data Paradigm
  - Big Data Science and Data Scientist as a new profession
  - Big Data Architecture

The **Big Data Paradigm** consists of the distribution of data systems across horizontally-coupled independent resources to achieve the scalability needed for the efficient processing of extensive datasets.
Main components of the Big Data ecosystem
- Data Provider
- Big Data Applications Provider
- Big Data Framework Provider
- Data Consumer
- Service Orchestrator

Big Data Lifecycle and Applications Provider activities
- Collection
- Preparation
- Analysis and Analytics
- Visualization
- Access

Big Data Ecosystem includes all components that are involved into Big Data production, processing, delivery, and consuming

(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 Infrastructure and Analytic Tools

Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.

Big Data Source/Origin (sensor, experiment, logdata, behavioral data)

Big Data Analytic/Tools
Analytics:
- Refinery, Linking, Fusion
- Realtime, Interactive, Batch, Streaming

Analytics Applications:
- Link Analysis
- Cluster Analysis
- Entity Resolution
- Complex Analysis

Analytics Applications:
- High Performance Computer Clusters
- Storage Specialised Databases
- Archives

Data Management
Data categories: metadata, (un)structured, (non)identifiable

Intercloud multi-provider heterogeneous Infrastructure
- Security Infrastructure
- Network Infrastructure Internal
- Infrastructure Management/Monitoring

Federated Access and Delivery Infrastructure (FADI)
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)
  - Intended to data centric, i.e. policy and access control are bound to data objects
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?
  - Traceability vs Opacity
  - Referral integrity
- Identifying and linking data
  - Persistent identifier

Common Data Model?
- Data Variety and Variability
- Semantic Interoperability
Evolutional/Hierarchical Data Model

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

- Referrals
- Control information
- Policy
- Data patterns

Usable Data

Actionable Data
Papers/Reports
Archival Data
Cloud HPC and Big Data Platforms

- **HPC on cloud platform**
  - Special HPC and GPU VM instances as well as Hadoop/HPC clusters offered by CSPs

- **Amazon Big Data services**
  - Amazon Elastic MapReduce

- **Microsoft Analytics Platform System (APS)**
  - Microsoft HD Insight

- **LexisNexis HPC Cluster System**
  - Combing both HPC cluster platform and optimized data processing languages

- **Streaming analytics/processing tools**
    - High throughput distributed subscription messaging system used for real-time stream processing
    - Distributed realtime computation system for unbounded streams of data
    - Is a new general-purpose fast cluster computing system that targets to combine batch and real-time streaming data processing
    - Spark provides primitives for in-memory cluster computing that makes it well suited to machine learning algorithms
Cloud and High Performance Computing (HPC)

- Current trends on increasing demand for HPC for Big Data Analytics and Business Intelligence
  - Data Analytics platforms and services are provided by the majority of big Cloud Service Providers as designated Data Analytics, Hadoop and HPC clusters
- Problem: Cloud Computing is not designed for HPC but rather for scalability and parallelization
  - Tightly coupled HPC workloads scale not well in a VM/cloud environment
- Restricting factors for HPC on cloud platform
  - High cost of data access/download from cloud
  - High-end VM configuration optimized for compute are proposed not by all clouds
- Computational challenges of Big Data can be solved by aggregating compute power and parallelisation of tasks
  - 24 hours on 1 machine can be 1 hour on 24 machines
  - These 24 machines require supporting infrastructure that will do: orchestration; jobs division into tasks, and distribution of tasks across a computer cluster.
AWS Cloud Big Data Services

AWS Cloud offers the following services and resources for Big Data processing:

• EC2 Virtual Machine (VM) instances for HPC optimized for computing (with multiple cores) and with extended storage for large data processing.

• **Amazon Kinesis** is a managed service for real-time processing of streaming big data (throughput scaling from megabytes to gigabytes of data per second and from hundreds of thousands different sources).

• **Amazon DynamoDB** highly scalable NoSQL data stores with sub-millisecond response latency.

• **Amazon Elastic MapReduce (EMR)** provides the Hadoop framework on Amazon EC2 and offers a wide range of Hadoop related tools.

• **Amazon Redshift** fully-managed petabyte-scale data warehouse in cloud at cost less than $1000 per terabyte per year. It is provided with columnar data storage with possibility to parallelise queries.

• Amazon RDS scalable relational database.

• Amazon Glacier archival storage to AWS for long time data storage at lower cost that standard Amazon S3 object storage.
Microsoft Azure Analytics Platform System (APS)

- Microsoft Azure cloud provides general IaaS services and reach Platform as a Service (PaaS) services.
  - Similar to AWS, Microsoft Azure offers special VM instances that have both computational and memory advanced capabilities.

- The Analytics Platform System (APS) combines the Microsoft SQL Server based Parallel Data Warehouse (PDW) platform with HDInsight and Apache Hadoop based scalable data analytics platform.
  - APS includes the PolyBase data querying technology to simplify integration of the PDW SQL data and data from Hadoop.

- HDInsight Hadoop based platform has been co-developed with Hortonworks
  - HDInsight provides comprehensive integration and management functionality for multi-workload data processing on Hadoop platform including batch, stream, in-memory processing methods.
HDInsight: Microsoft’s Big Data Solution

- HDInsight can run both on Azure Cloud and on Windows Server (on premises)
  - Data exchange via PolyBase
- Compatible with and support all products from Apache Hadoop stack
- Supports all stages of Big Data processing
- PolyBase is a new technology that allows integrating Microsoft SQL Server based Parallel Data Warehouse (PDW) with Hadoop
- Azure Blob Storage used to persistently store data
  - Data are streamed to Hadoop/HDFS for processing and pushed back to Azure Blob Storage
HDInsight/Hadoop Ecosystem

Legend

Red = Core Hadoop

Blue = Data processing

Purple = Microsoft integration points and enhancements

Orange = Data Movement

Green = Packages

[ref] Microsoft Azure Training Kit. [https://github.com/Azure-Readiness/MicrosoftAzureTrainingKit](https://github.com/Azure-Readiness/MicrosoftAzureTrainingKit)
LexisNexis HPCC Systems as an integrated Open Source platform for Big Data Analytics

HPCC Systems data analytics environment components and HPCC Systems architecture model is based on a distributed, shared-nothing architecture and contains two clusters:

- **THOR Data Refinery**: THOR is a massively parallel Extract, Transform, and Load (ECL) engine that can be used for performing a variety of tasks such as massive joins, merges, sorts, transformations, clustering, and scaling.
- **ROXIE Data Delivery**: ROXIE serves as a massively parallel, high throughput, structured query response engine.
  - Perform volumes of structured queries and full text ranked Boolean search.
  - ROXIE also provides real-time analytics capabilities, to address real-time classifications, prediction, fraud detection and other problems that normally require stream analytics.

Other components of the HPCC environment

- **Enterprise Control Language (ECL)**: An open source, data-centric declarative programming language used by both THOR and ROXIE for large-scale data management and query processing.
- **ECL compiler and job server (ECL Server)**: Serves as the code generator and compiler that translates ECL code.
- **System data store (Dali)**: Used for environment configuration, message queue maintenance, and enforcement of LDAP security restrictions.
- **Archiving server (Sasha)**: Serves as a companion ‘housekeeping’ server to Dali.
- **Distributed File Utility (DFU Server)**: Controls the spraying and despraying operations used to move data onto and out of THOR.
- **The inter-component communication server (ESP Server)**: Supports multi-protocol communication between services to enable various types of functionality to client applications via multiple protocols.
LexisNexis HPCC Systems Architecture

- THOR is used for massive data processing in batch mode for ETL processing
- ROXIE is used for massive query processing and real-time analytics
LexisNexis HPCC Systems Data Analytics Environment

- Unstructured and structured content
  - Ingest disparate data sources
  - Manage hundreds of TB of data
- Data Refinery, Linking Fusion
  - Parallel Processing Architecture
  - Discover non-obvious links and detect hidden patterns
  - Language optimised for data-intensive Application
- Analysis Applications
  - Entity resolution
  - Link Analysis
  - Clustering Analysis
  - Complex Analysis
- Variety of industry applications

IEEE Cloud Course
Cloud and Big Data
LexisNexis Data Analytics Languages

- **Enterprise Control Language (ECL):** An open source, data-centric declarative programming language
  - The declarative character of ECL language simplifies coding
  - ECL is developed to simplify both data query design and customary data transformation programming
  - ECL is explicitly parallel and relies on the platform parallelism.
  - *x100 less code than SQL*

- LexisNexis proprietary *record linkage technology SALT (Scalable Automated Linking Technology)* that automates data preparation process: profiling, parsing, cleansing, normalisation, standardisation of data.
  - SALT uses weighted matching and threshold based computation, it also enables internal, external and remote linking with external or master datasets.
  - Enables the power of the HPCC Systems and ECL

- **Knowledge Engineering Language (KEL) is an ongoing development**
  - KEL is a domain specific data processing language that allows using semantic relations between entities to automate generation of ECL code.
  - *x10 less code than ECL*

- The report is the result of the 90-day study commissioned by the President of the United States to examine how big data will transform the way people live and work and how big data will alter the relationships between government, citizens, businesses and consumers.

Data security and privacy challenges in Cloud Computing and big data have been a focus of numerous study groups initiated by different governmental bodies that produced several valuable reports.

- Wide implementation of Cloud Computing provided a basis for developing big data technologies and data-centric and data-driven applications that in their own turn facilitate cloud technologies development.

The main approach in developing recommendations was to protect privacy while not hindering/restricting development of new technology for the benefit of the whole society.

- The report expresses the opinion that despite widely discussed needs for personal control of the collected e-commerce and social data, the practical use of such control is impractical due to the unmanageable volume of information and its variety. Instead, the advertisement companies and other organisational users of the personally...
A number of recent documents developed by President Obama’s administration provide a foundation for the Right for Privacy in our Information society. A key document in this respect is the ‘Consumer Data Privacy In A Networked World: A Framework For Protecting Privacy And Promoting Innovation In The Global Digital Economy’ published in February 2012, which states the following rights:

- **Individual Control**: Consumers have a right to exercise control over what personal data organisations collect from them and how they use it.
- **Transparency**: Consumers have a right to easily understandable information about privacy and security practices.
- **Respect for Context**: Consumers have a right to expect that organisations will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- **Security**: Consumers have a right to secure and responsible handling of personal data.
- **Access and Accuracy**: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data are inaccurate.
- **Focused Collection**: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
- **Accountability**: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.

The theory behind the Consumer Privacy Bill of Rights and other documents is that the combination of broad baseline principles and specific codes of conduct can protect consumers while supporting innovation.
EU documents outlining EU regulatory basis for data protection in cloud

- Protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 25.1.2012, COM(2012)

- Everyone has the right to the protection of personal data
  - Under EU law, personal data can only be gathered legally under strict conditions, for a legitimate purpose.
  - Furthermore, persons or organisations which collect and manage your personal information must protect it from misuse and must respect certain rights of the data owners which are guaranteed by EU law.

- Privacy impact and requirements
  - Right to be forgotten (RTBF) – complex issue for global cloud infrastructures and all information collected on the web and mobile applications

- Recommendation to establish a position responsible for compliance
- Privacy by-design principle for services and infrastructure
In order to bridge differences in approach and provide a streamlined means for U.S. organizations to comply with the EU Directive, the U.S. Department of Commerce in consultation with the European Commission developed a so-called "Safe Harbor" framework that defines key principles and provides recommendations for companies to protect personal data and personal information.

• "Personal data" and "personal information" are data about an identified or identifiable individual that are recorded in any form.

7 Safe Harbor privacy principles

• NOTICE: An organization must inform individuals about the purposes for which it collects and uses information about them, the types of third parties to which it discloses the information, and the choices and means the organization offers individuals for limiting its use and disclosure.

• CHOICE: An organization must offer individuals the opportunity to choose (opt out) whether their personal information is (a) to be disclosed to a third party, or (b) to be used for a purpose that is incompatible with the purpose(s) for which it was originally collected.

• ONWARD TRANSFER: To disclose information to a third party, organizations must apply the Notice and Choice Principles.

• SECURITY: Organizations creating, maintaining, using or disseminating personal information must take precautions to protect it from loss, misuse and unauthorized access, disclosure, alteration and destruction.

• DATA INTEGRITY: Consistent with the Principles, personal information must be relevant for the purposes for which it is to be used.

• ACCESS: Individuals must have access to personal information about them that an organization holds and be able to correct, amend, or delete.

• ENFORCEMENT: Effective privacy protection must include mechanisms for assuring compliance with the Principles, and consequences for the organization when the Principles are not followed.
• Your possible target for 2-4 years
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
### 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

<table>
<thead>
<tr>
<th>Developer</th>
<th>Engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Researcher</td>
<td>Data Creative</td>
</tr>
<tr>
<td>Data Businessperson</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>Big and Distributed Data</td>
<td>Structured Data</td>
</tr>
<tr>
<td>Unstructured Data</td>
<td>Optimization</td>
</tr>
<tr>
<td>Product Development</td>
<td>Systems Administration</td>
</tr>
<tr>
<td>Business</td>
<td>Back End Programming</td>
</tr>
</tbody>
</table>

#### Statistics

- Visualization
- Temporal Statistics
- Surveys and Marketing
- Spatial Statistics
- Data Manipulation
- Classical Statistics

#### Business

- Back End Programming
- Systems Administration
- Business

#### ML / Big Data

- Big and Distributed Data
- Structured Data
- Machine Learning
- Unstructured Data

#### Math / OR

- Algorithms
- Graphical Models
- Bayesian / Monte Carlo Statistics
-Simulation

#### Programming

- Front End Programming
Skills and Self-ID Top Factors

- Business
- ML/BigData
- Math/OR
- Programming
- Statistics

ML – Machine Learning
OR – Operations Research

2014, UvA
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**
Questions and Discussion

• Master Research Projects
  – Big Data Infrastructure and Tools + Data protection
  – Cloud Computing + Cloud Security
  – Advanced Networking for Cloud infrastructures
Common Body of Knowledge (CBK) in Cloud Computing

CBK refers to several domains or operational categories into which Cloud Computing theory and practices breaks down
- Creates a basis for education and training
- Defines qualification and knowledge matrix/vocabulary

CBK Cloud Computing elements (basic and extended)
1. Cloud Computing Architectures, service and deployment models
2. Cloud Computing platforms, software/middleware and API’s
3. Cloud Services Engineering, Cloud aware Services Design
4. Virtualisation technologies (Compute, Storage, Network)
5. Computer Networks, Software Defined Networks (SDN)
6. Service Computing, Web Services and Service Oriented Architecture (SOA)
7. Computing models: Grid, Distributed, Cluster Computing
8. Security Architecture and Models, Operational Security
9. IT Service Management, Business Continuity Planning (BCP)
10. Business and Operational Models, Compliance, Assurance, Certification

Responsibilities Split in IaaS, PaaS, SaaS

Security management responsibilities split between Customer and Provider for IaaS, PaaS, SaaS service models:

- Updating firmware and software for platform and for customer managed components
- Firewall is intrusion prevention is a responsibility of the cloud provider
- Certification and compliance of the cloud platform doesn’t imply security and compliance of the customer controlled components
For other cloud service models PaaS and SaaS the responsibility of AWS goes up to OS, network and firewall for PaaS, and also includes the application platform and container for SaaS.

- However, the responsibility for data remains with the customer.