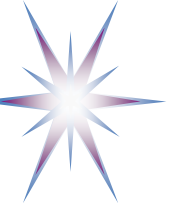


Cloud based Big Data Platforms and New Profession of Data Scientist

Konferencję Użytkowników Komputerów Dużej Mocy – KU KDM'16
17 March 2016, Zakopane, Poland

Yuri Demchenko, University of Amsterdam

Outline

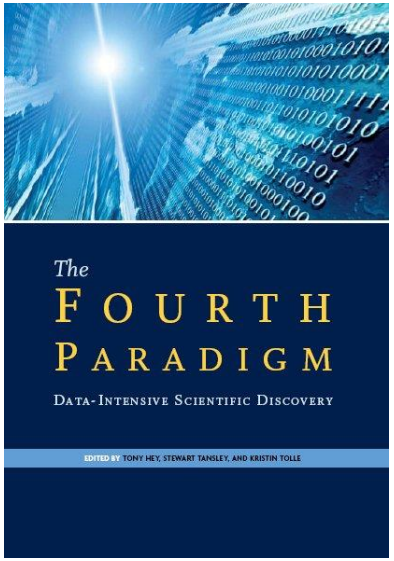
- Big Data and Data Centric Computing
  - Need for new paradigms, architecture and platforms
- Cloud Computing as a platform of choice for Big Data applications
  - Big Data Stack and cloud advantages
  - Cloud platforms for Big Data
- Big Data, Data Science and Data Scientist profession definition
  - Data Science competences, skills and body of knowledge

Acknowledgement
Slides on Big Data Stacks are credit to David Bernstein, Cloud Strategy Partners

Yuri Demchenko – Professional Summary

- Graduated from National Technical University of Ukraine “Kiev Polytechnic Institute” (KPI) in Instrumentation and Measurement (aka Industry Automation)
  - Candidate of Science (Tech) – Dissertation on System Oriented Precision Generators (1989)
- Teaching at KPI 1989-1998 – Computer Networking, Internet Technologies, Security
- Professional work in Internet technologies since 1993
- Work at TERENA (Trans-European R&E Networking Association) – 1998-2002
- Work at UvA with SNE group – since 2003
  - Main research areas: Cloud Computing, Big Data Infrastructures, Application and Infrastructure Security, Generic AAA&Authorisation, Grid and collaborative systems
  - EU Projects: GEYSERS, GEANT3, Phosphorus, EGEE I-II, Collaboratory.nl
  - Standardisation activity – IETF, Open Grid Forum (OGF) – ISOD-RG chairing, NIST Cloud Collaboration, NIST Big Data WG, ISO/IEC Big Data Study Group
  - Now/2014: Big Data Architecture, Big Data Security, Cloud Computing and Big Data Curriculum development
  - Now/2015: EDSION Project coordinator: Building Data Science Profession

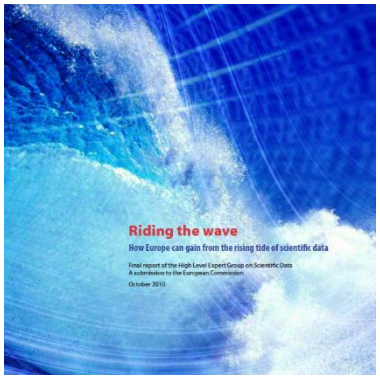
Visionaries and Drivers: Seminal works, High level reports, Activities



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)
http://bigdatawg.nist.gov/

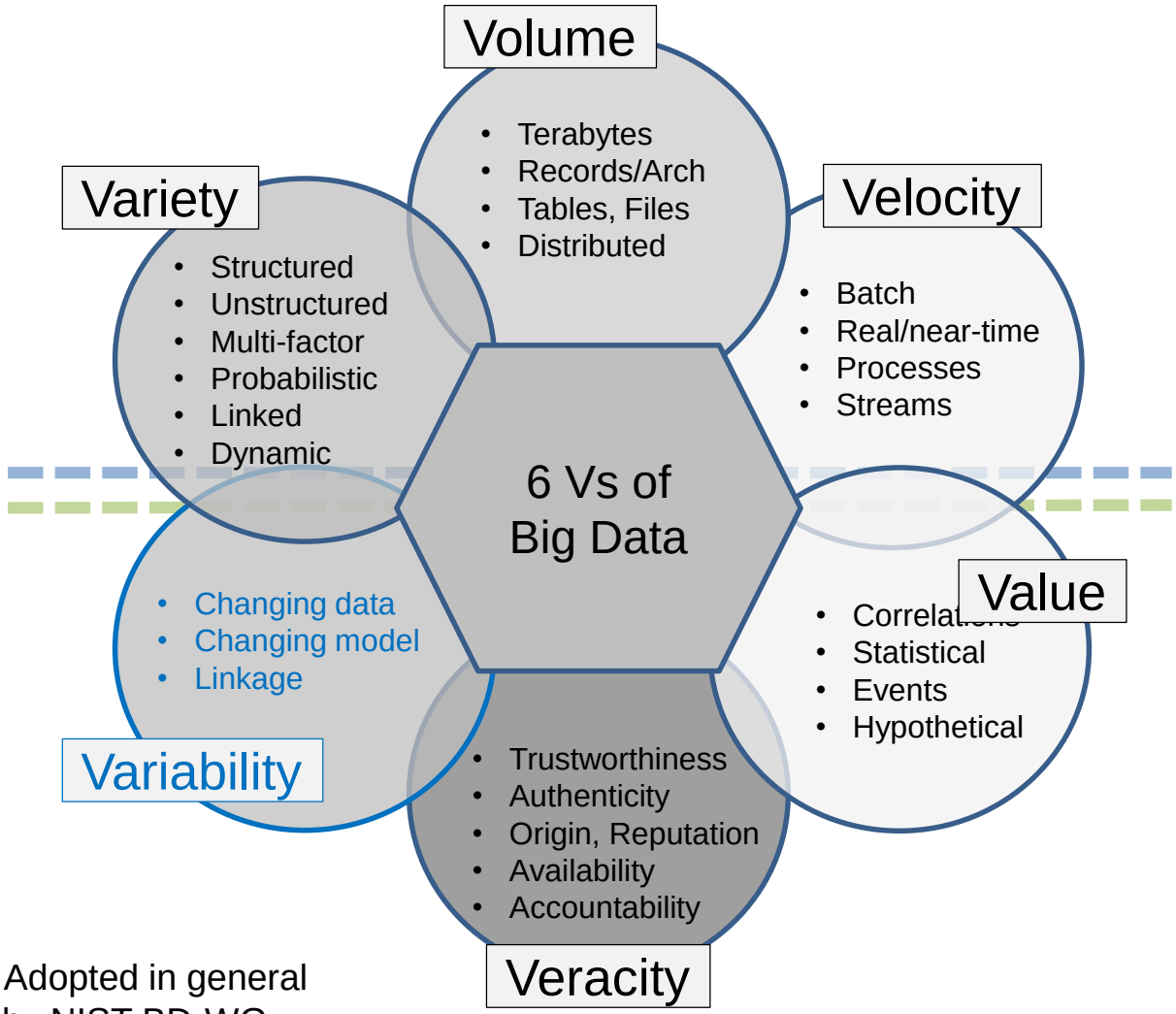
ISO/IEC JTC1 Big Data Study Group (SGBD)
http://jtc1bigdatasg.nist.gov/home.php

The Data Harvest: How sharing research data can yield knowledge, jobs and growth.

An RDA Europe Report. December 2014

https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

Big Data definition revisited: 6 V's of Big Data



Adopted in general by NIST BD-WG

Generic Big Data Properties

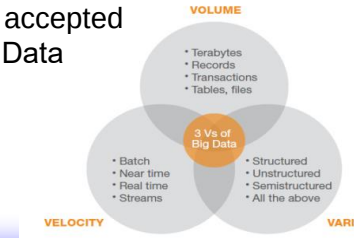
- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data

- Volume
- Velocity
- Variety



Big Data definition revisited: 5 parts vs 6V

(1) Big Data Properties: 6V
- 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

NIST Big Data Working Group (NBD-WG) and ISO/IEC JTC1 Study Group on Big Data (SGBD)

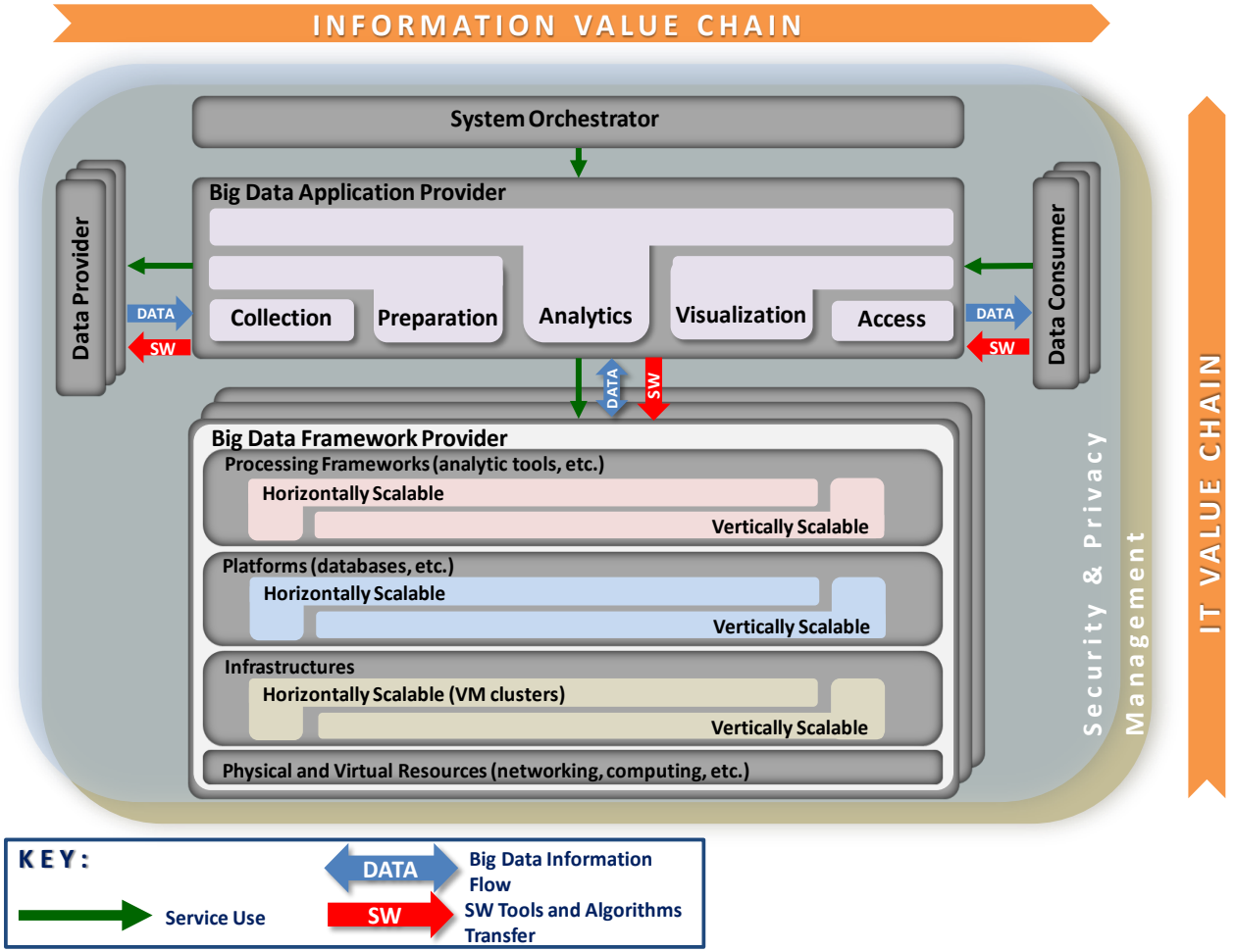
- NIST Big Data Working Group (NBD-WG) is leading the development of the Big Data Technology Roadmap - http://bigdatawg.nist.gov/home.php
  - Built on experience of developing the Cloud Computing standards fully accepted by industry
- Set of documents published in September 2015 as NIST Special Publication NIST SP 1500: NIST Big Data Interoperability Framework (NBDIF) http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

  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.

NIST Big Data Reference Architecture



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

[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

Big Data Architecture Framework (BDAF) by UvA

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

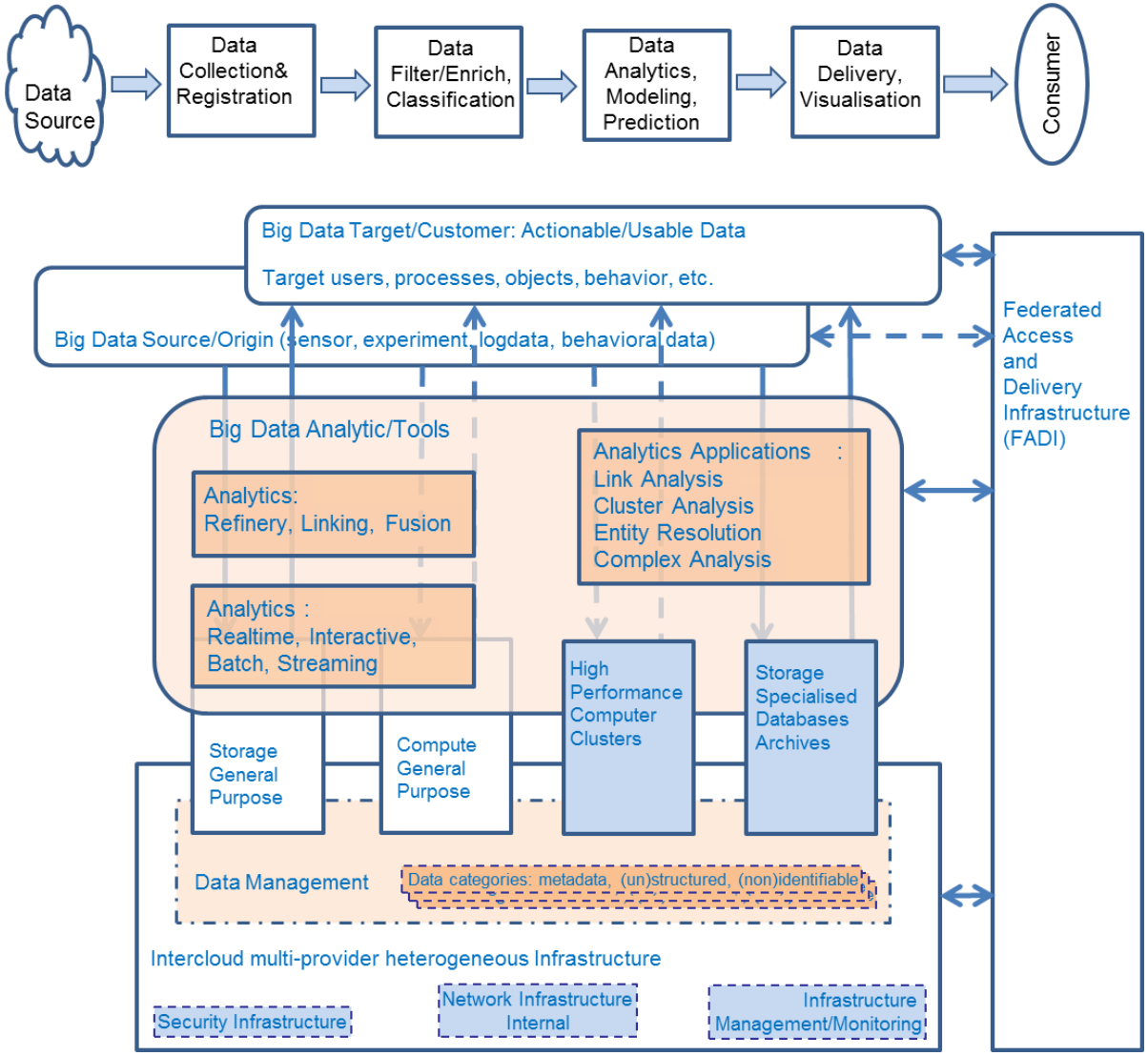
(2) Big Data Management
- Big Data Lifecycle (Management) Model
  - Big Data transformation/staging
- Provenance, Curation, Archiving

(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 Analytics Tools



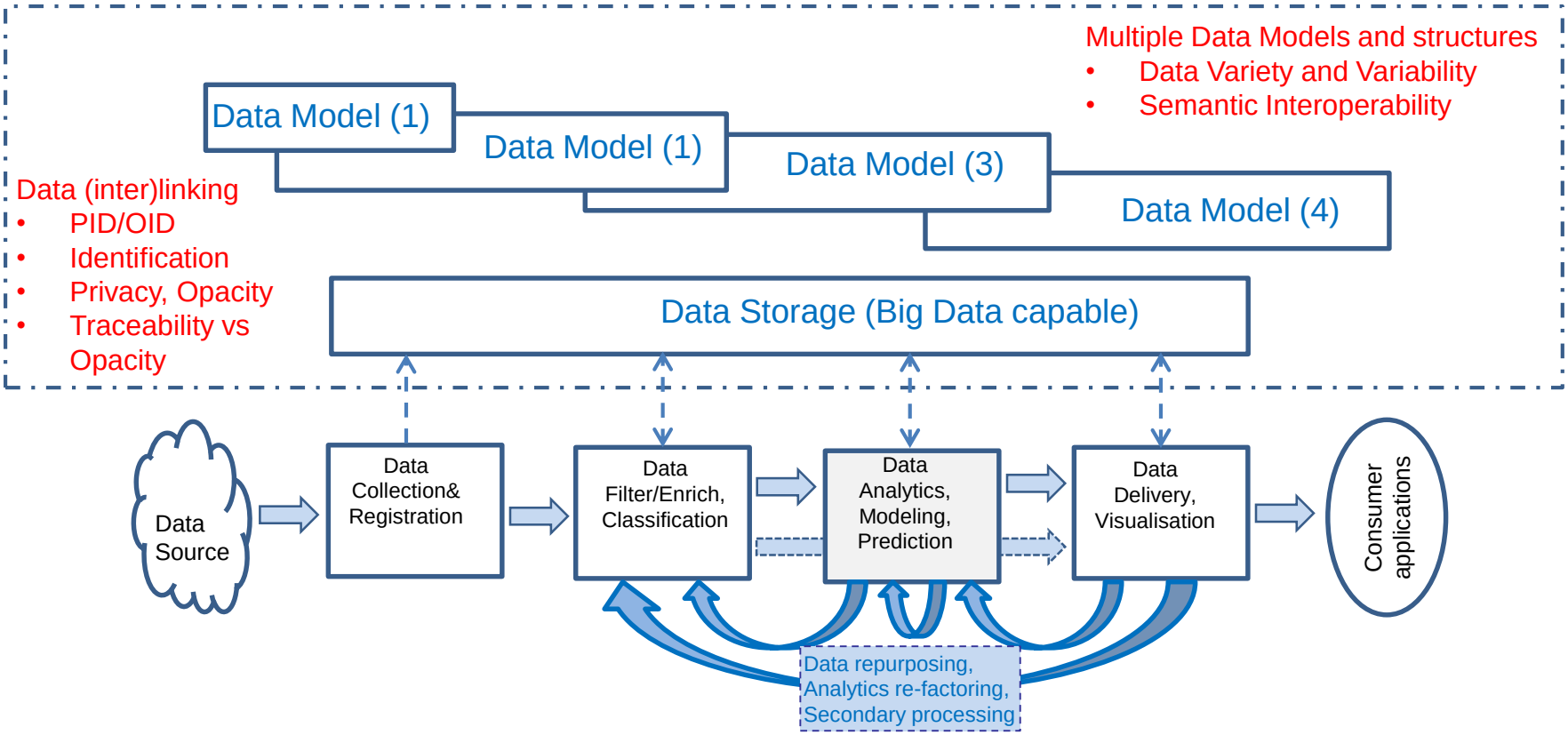
Big Data Infrastructure

- Heterogeneous multi-provider inter-cloud infrastructure
- Data management infrastructure
- Collaborative Environment
- Advanced high performance (programmable) network
- Security infrastructure
- Federated Access and Delivery Infrastructure (FADI)

Big Data Analytics Infrastructure/Tools

- High Performance Computer Clusters (HPCC)
- Big Data storage and databases SQL and NoSQL
- Analytics/processing: Real-time, Interactive, Batch, Streaming
- Big Data Analytics tools and applications

Data Lifecycle/Transformation Model



- Data Model changes along data lifecycle or evolution
- Data provenance is a discipline to track all data transformations along lifecycle

- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity

Big Data is Driving Cloud Usage – Cloud powers Big Data applications

Supply

There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing

Eric Schmidt, Google CEO, Techonomy Conference, August 4, 2010

LARGE

Real-time

Data is becoming the new raw material of business: an economic input almost on a par with capital and labour. "Every day I wake up and ask, 'how can I flow data better, manage data better, analyse data better?" says Rollin Ford, the CIO of Wal-Mart.

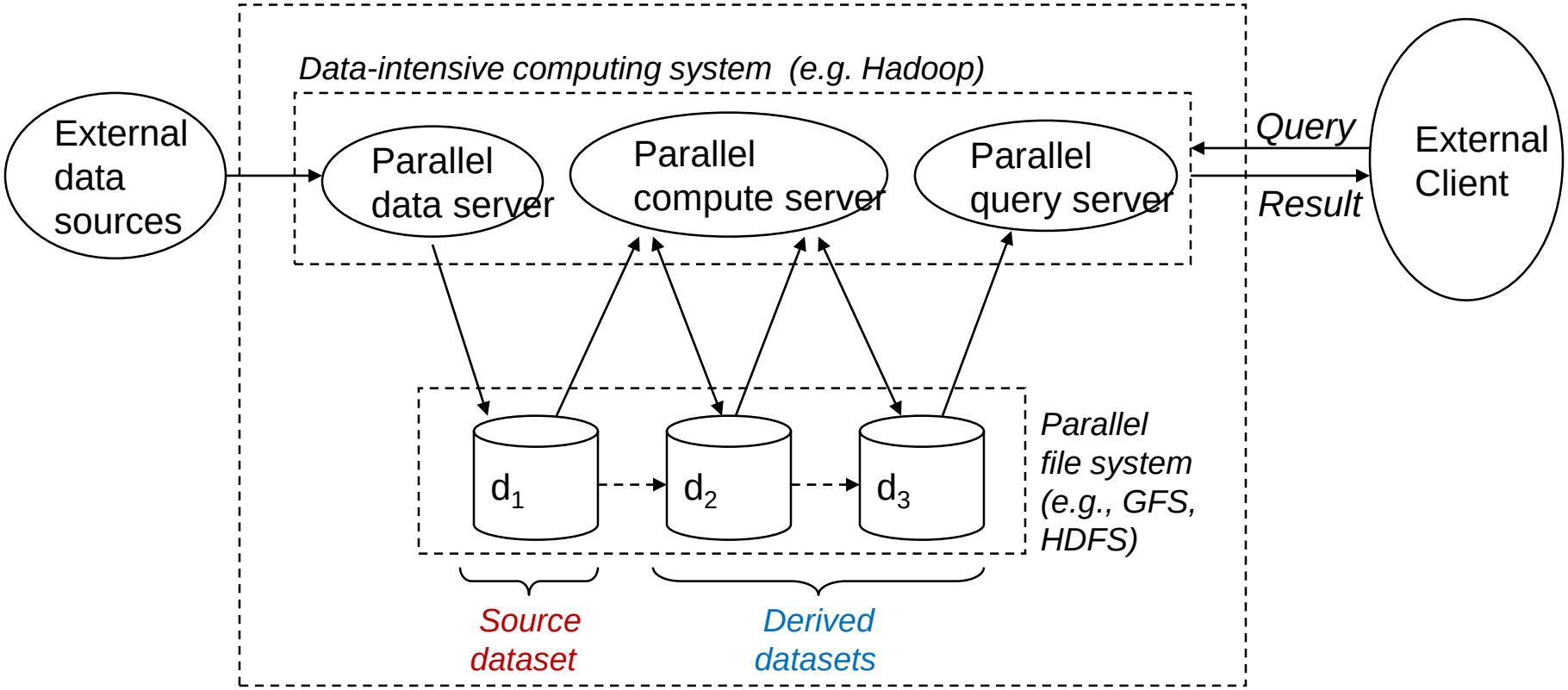
Source: Data, Data Everywhere, The Economist, February 25, 2010

Demand

Public Clouds are All Over the Place

Cloud Centers are All Over the Place, from datacentermap.com November 2015
http://www.datacentermap.com/cloud.html

Cloud Based Big Data Services



Characteristics:
Massive data and computation on cloud, small queries and results

Examples:
Search, scene completion service, log processing

Big Data Stack

Hook into an existing queue to get copy / subset of data. Queue handles partitioning, replication, and ordering of data, can manage backpressure from slower downstream components

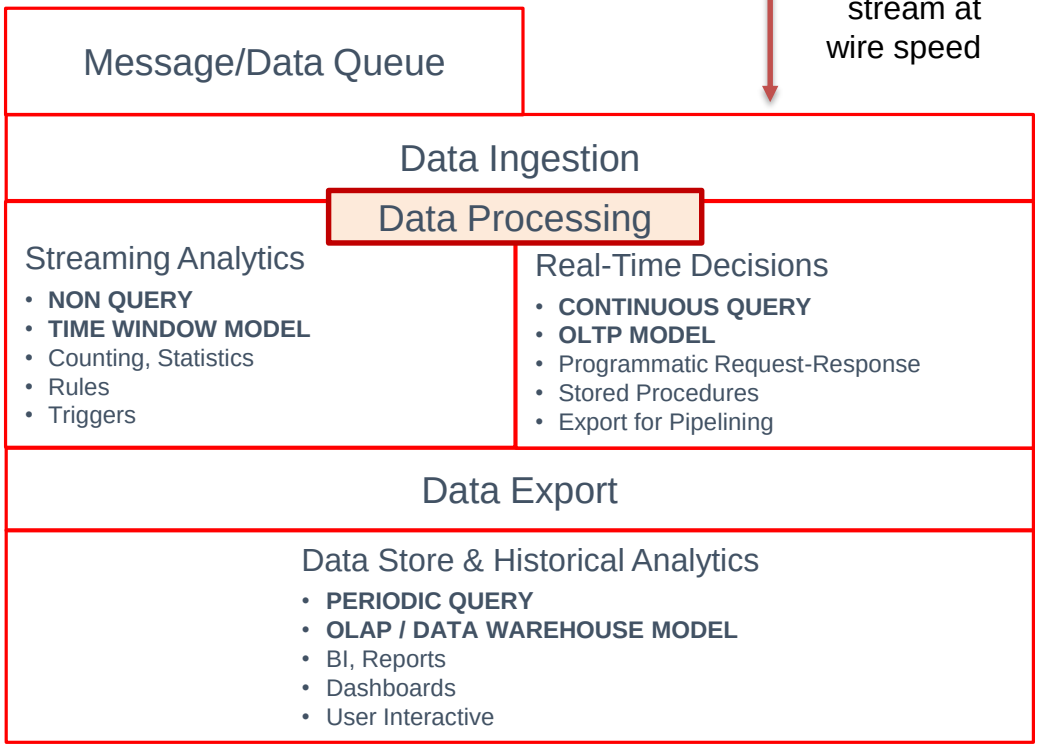
Use direct Ingestion API to capture entire stream at wire speed

Real-Time Decisions and/or Results can be fed back "up-stream" to influence the "next step"

Ingestion will transform, normalize, distribute and integrate to one or more of the Analytic / Decision Engines of choice

Use one or more Analytic / Decision engines to accomplish specific task on the Fast Data window

The OLAP Engines of choice will append to historical data and store for later, further Big Data analysis on entire data set



Export will transform, normalize, distribute and integrate to one or more of the Data Warehouse / Storage Engines of choice

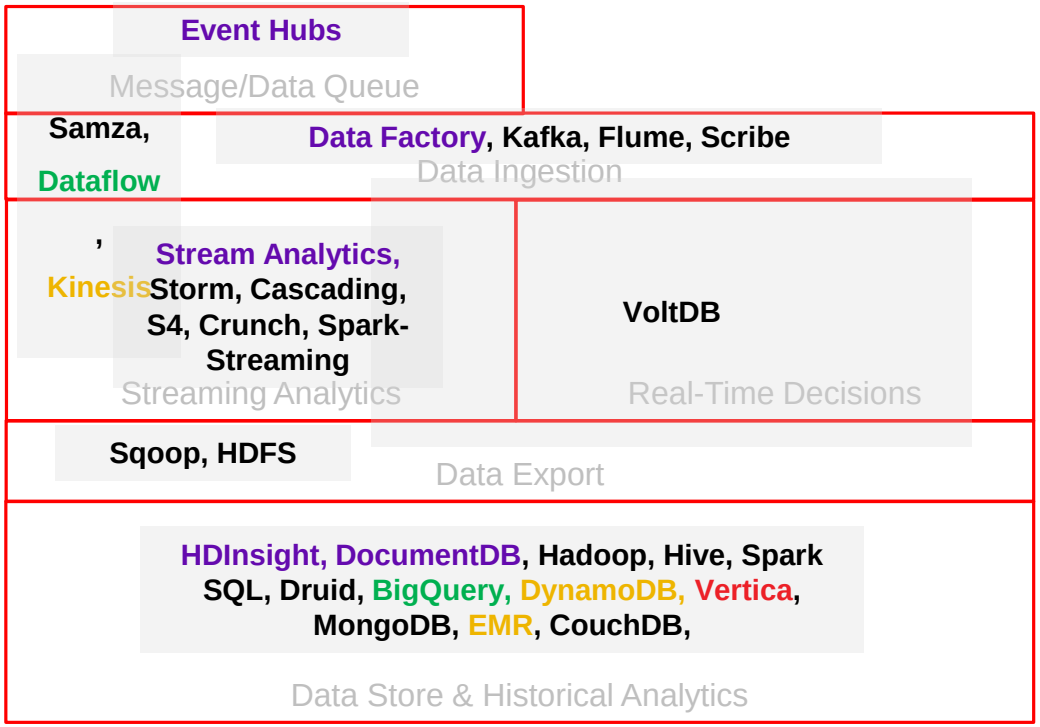
Important Big Data Technologies

Microsoft Azure:
Event Hubs
Data Factory
Stream Analytics
HDInsight
DocumentDB

Google GCE:
DataFlow
BigQuery

Amazon AWS:
Kinesis
EMR
DynamoDB

Proprietary:
Vertica
HortonWorks

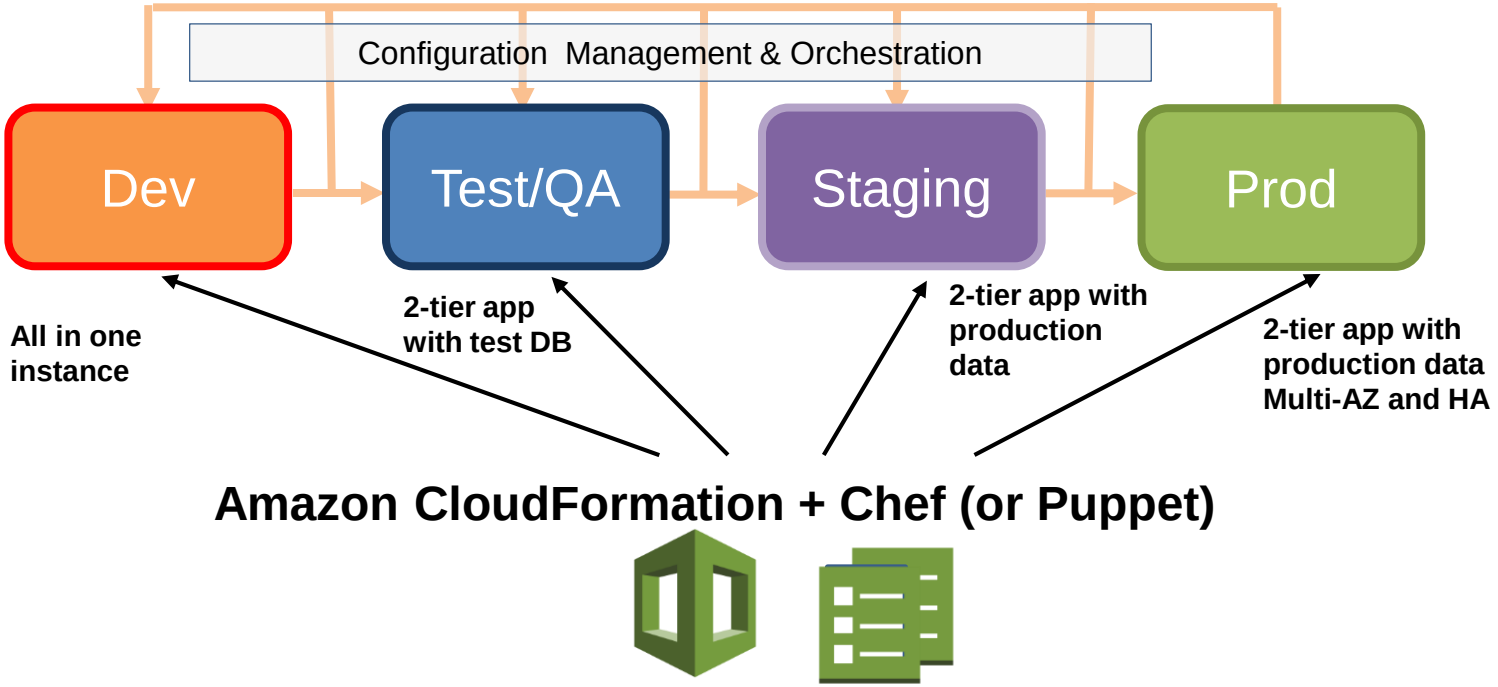


Open Source
Samza
Kafka
Flume
Scribe
Storm
Cascading
S4
Crunch
Spark
Hadoop
Hive
Druid
MongoDB
CouchDB
VoltDB

Cloud Platform Benefits for Big Data

- Segregated networks isolate traffic
  - Clouds construction provides separate networks for each type of traffic
  - Big Data applications benefit from lowest latencies possible for node to node synchronization, dynamic cluster resizing, and other scale-out operations
- Cloud deployment on virtual machines, containers, and bare metal
  - For a traditional highly load-variable problem one might consider using VM's as a deployment vehicle for that.
  - Some clouds will offer container based isolation instead of VMs.
- Cloud tools for large scale applications deployment and automation
  - Supported by major IDE
  - Basis for agile technologies and Zero-touch services provisioning

Cloud-powered Services Development Lifecycle: DevOps == Continuous service improvement



- Easily creates test environment close to real
- Powered by cloud deployment automation tools
  - To enable configuration Management and Orchestration, Deployment automation
- Continuous development – test – integration
  - CloudFormation Template, Configuration Template, Bootstrap Template
- Can be used with Puppet and Chef, two configuration and deployment management systems for clouds

[ref] 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 HPC and Big Data Platforms

- HPC on cloud platform
  - Special HPC and GPU VM instances as well as Hadoop/HPC clusters offered by all CSPs
- Amazon Big Data services
  - Amazon Elastic MapReduce, Kinesis, DynamoDB, Regshift, etc
- Microsoft Analytics Platform System (APS)
  - Microsoft HD Insight/Hadoop ecosystems
- IBM BlueMix applications development platform
  - Includes full cloud services and data analytics services
- LexisNexis HPC Cluster System
  - Combing both HPC cluster platform and optimized data processing languages
- Variety of Open Source tools
  - Streaming analytics/processing tools: Apache Kafka, Apache Storm, Apache Spark

AWS Cloud Big Data Services

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

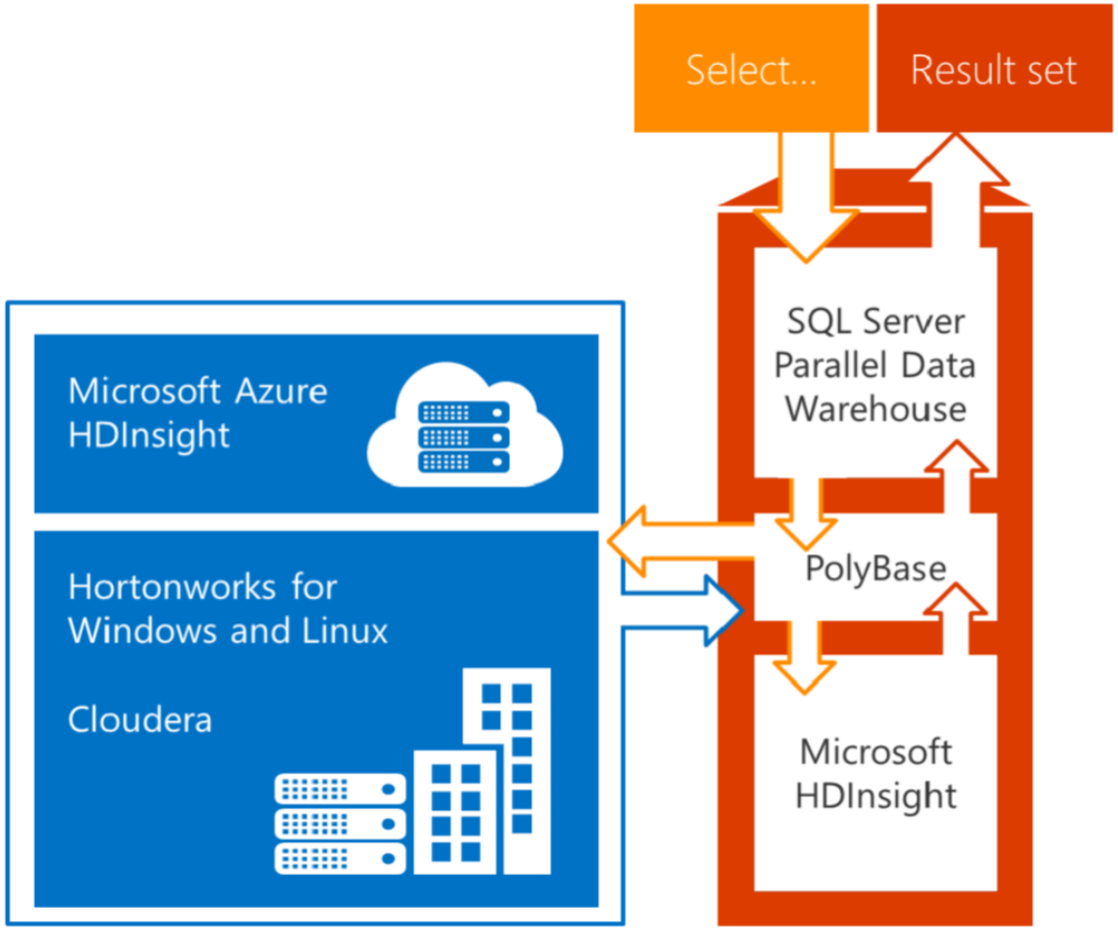
- EC2 Virtual Machine (VM) instances for HPC optimized for computing (with multiple cores) and with extended storage for large data processing.
- **Amazon Elastic MapReduce (EMR)** provides the Hadoop framework on Amazon EC2 and offers a wide range of Hadoop related tools.
- **Amazon Kinesis** is a managed 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 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.

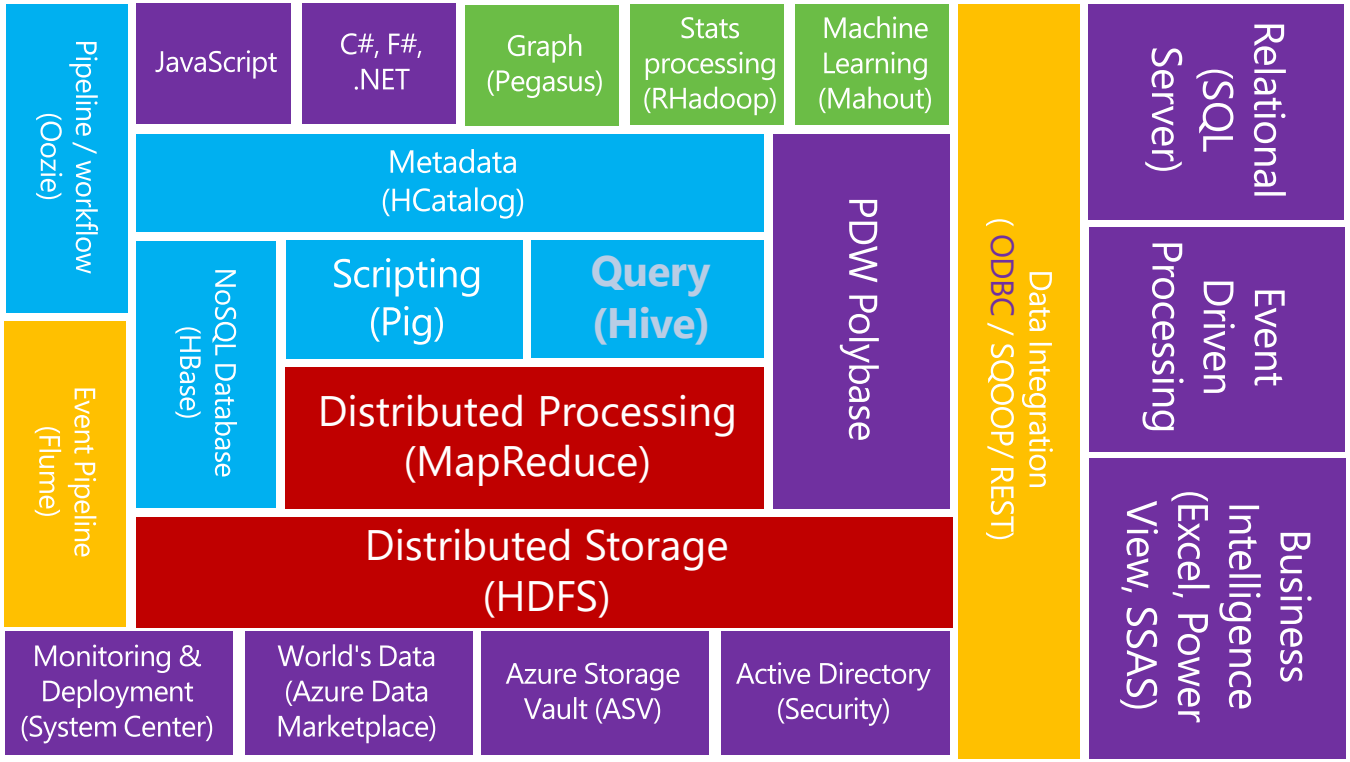
- The Analytics Platform System (APS) combines the Microsoft SQL Server based Parallel Data Warehouse (PDW) platform with HDInsight and Apache Hadoop based scalable data analytics platform.
  - APS includes the PolyBase data querying technology to simplify integration of the PDW SQL data and data from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

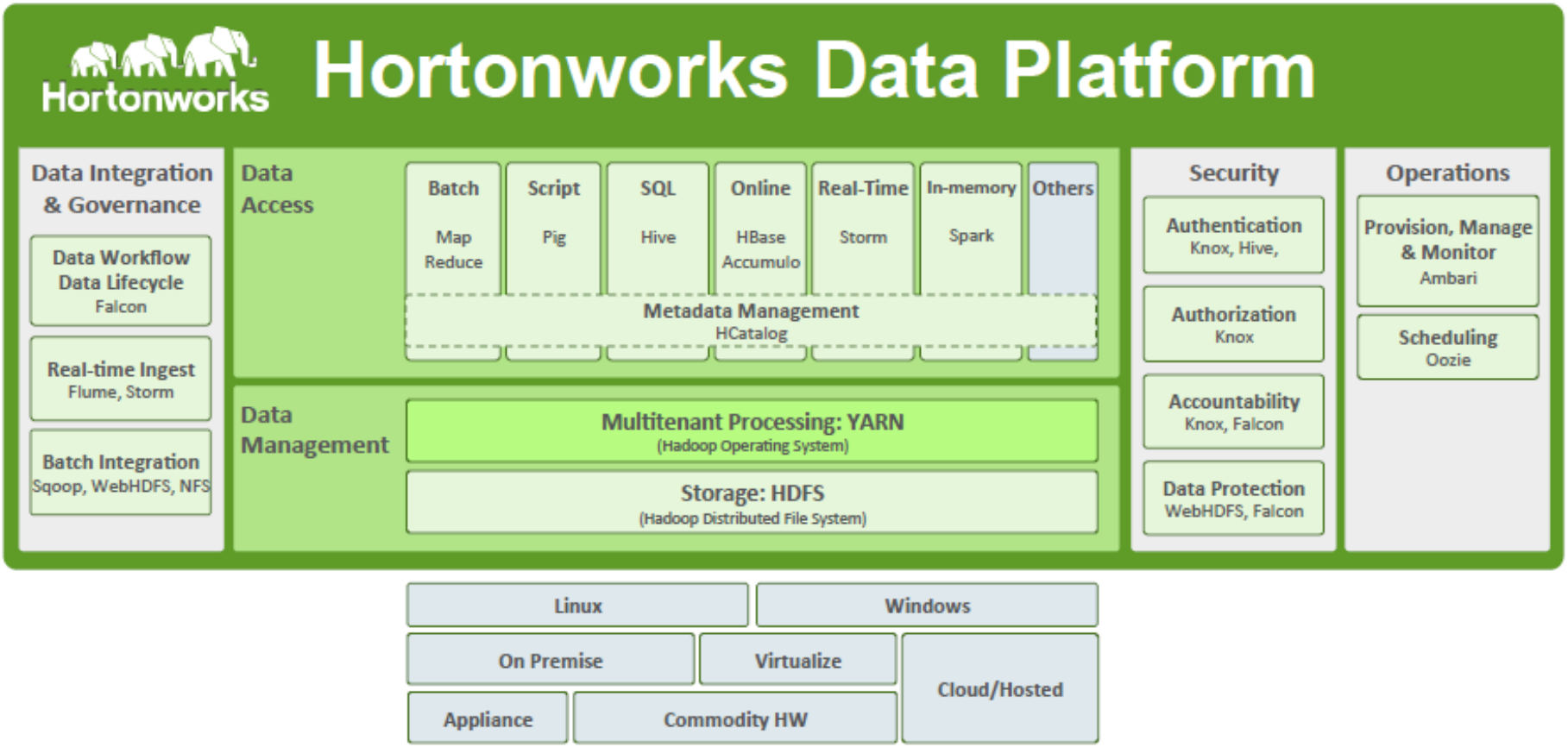
Purple = Microsoft integration points and enhancements

Orange = Data Movement

Green = Packages

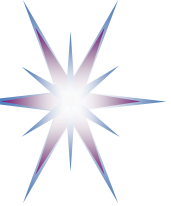
[ref] Microsoft Azure Training Kit. https://github.com/Azure-Readiness/MicrosoftAzureTrainingKit

Hortonworks Data Platform Architecture [ref]



- HDP includes the most recent developments of the Open Source Hadoop suite
- Can run on Linux and on Windows OS
- Can be deployed on premises on dedicated cluster and on cloud as a hosted application

[ref] http://hortonworks.com/hdp /



Hortonworks Data Platform (HDP) http://hortonworks.com/

- HDP delivers a single integrated Hadoop platform for enterprises
  - Provides a data platform for multi-workload data processing across an array of processing methods including batch and interactive to real-time
  - Supports key capabilities of an enterprise data platform: Governance, Security and Operations
  - YARN and Hadoop Distributed Filesystem (HDFS) are the core components of HDP
- YARN is treated as datacenter OS and supports multiple access methods (batch, real-time, streaming, in-memory, and more) on a common data set
  - YARN is the architectural center of Hadoop that allows to process data simultaneously in multiple ways
  - Allows creating multi-tenant data analytics applications
- HDP runs natively on Linux and Windows OS
  - HDP provides the basis for Microsoft's HDInsight Service meaning complete portability of data is retained on-premise and in the cloud
  - Available in integrated hardware from Teradata
- Hortonworks provides a simple starters solution Hadoop Sandbox
  - Hortonworks Sandbox is a single-node implementation of Hadoop based on the Hortonworks Data Platform that includes all the typical components found in a Hadoop deployment

LexisNexis HPCC Systems as an integrated Open Source platform for Big Data Analytics

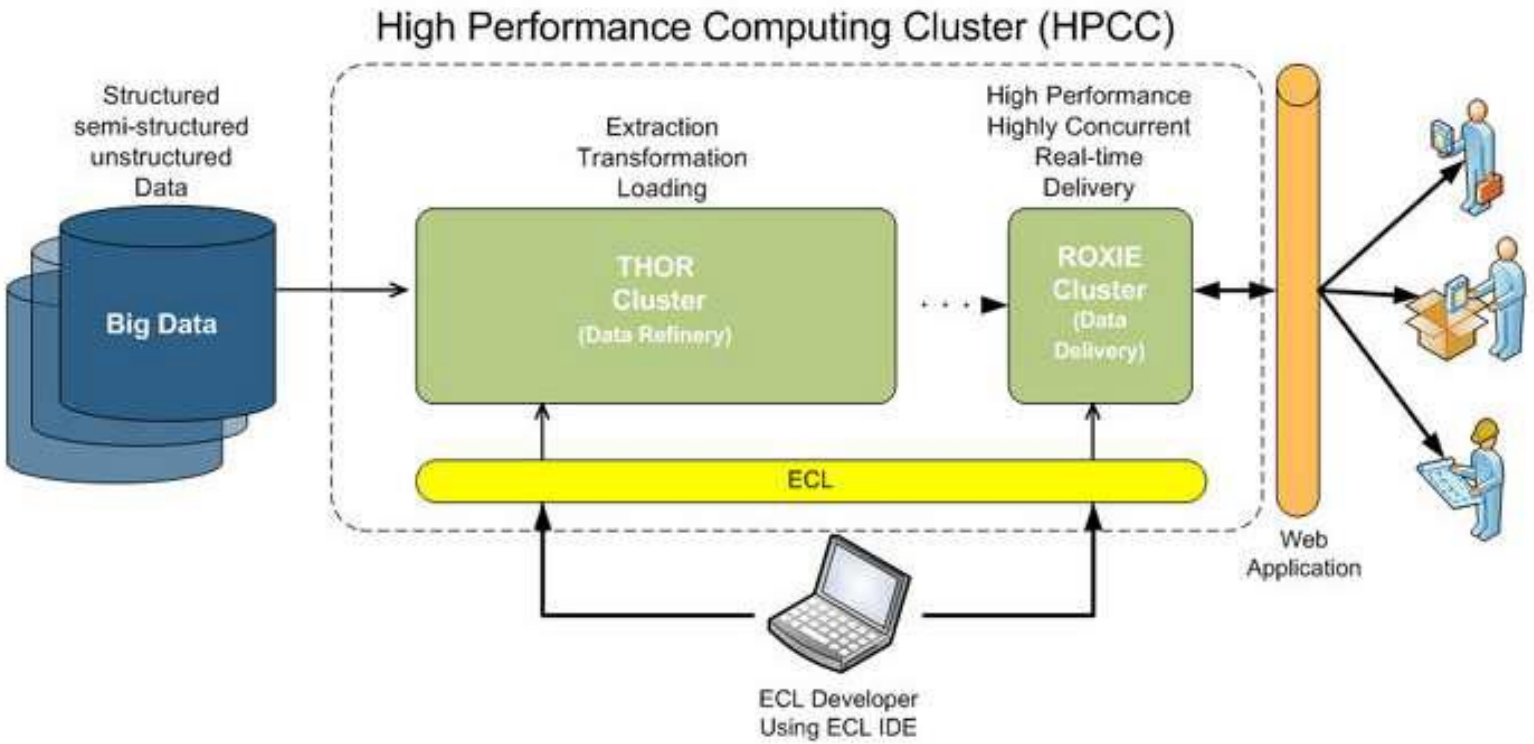
HPCC Systems data analytics environment components and HPCC Systems architecture model is based on a distributed, shared-nothing architecture and contains two cluster

- **THOR Data Refinery**: Massively parallel Extract, Transform, and Load (ECL) engine that can be used for variety of tasks such as massive: joins, merges, sorts, transformations, clustering, and scaling.
- **ROXIE Data Delivery**: Massively parallel, high throughput, structured query response engine with real time analytics capability

Other components of the HPCC environment: 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 explicitly parallel and relies on the platform parallelism.
- LexisNexis proprietary record linkage technology **SALT (Scalable Automated Linking Technology)**: automates data preparation process: profiling, parsing, cleansing, normalisation, standardisation of data.
  - 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.

LexisNexis HPCC Systems Architecture



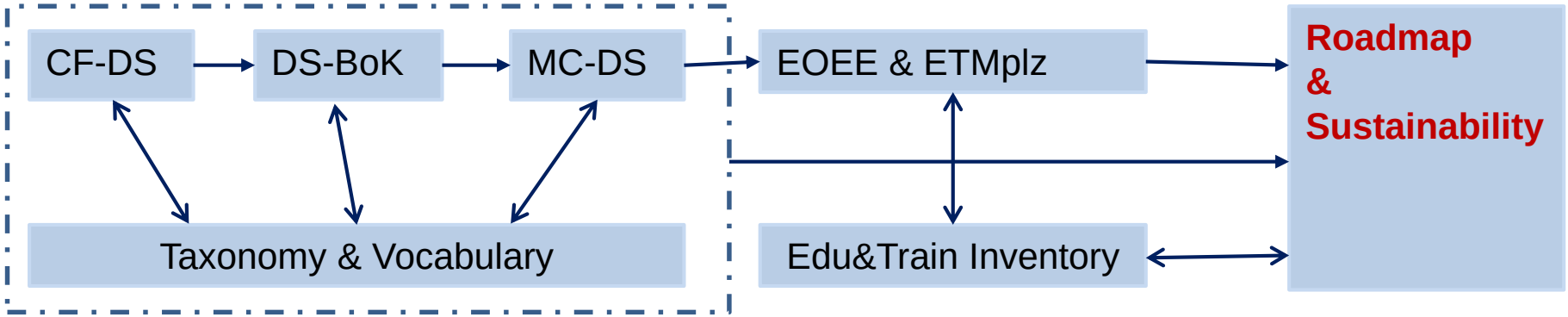
- THOR is used for massive data processing in batch mode for ETL processing
- ROXIE is used for massive query processing and real-time analytics

Data Science Profession Definition

New technologies require new competences, skills and new professions

- EDISON Data Science Framework
- Data Science professions family
- EU activities to address e-skills shortage
- EDISON engagement and outreach activities

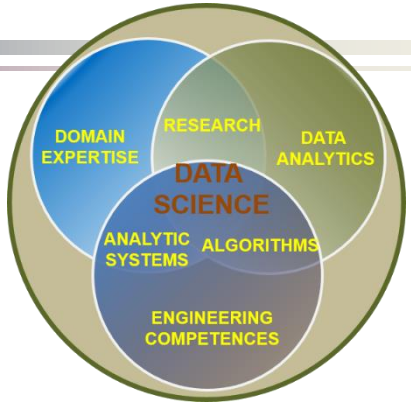
EDISON Framework and Background Developments



- EDISON Framework components
  - CF-DS – Data Science Competence Framework
  - DS-BoK – Data Science Body of Knowledge
  - MC-DS – Data Science Model Curriculum
  - Data Science Taxonomy and Scientific Disciplines Classification
  - EOEE - EDISON Online Education Environment

- Background: EU Competence Frameworks and Profiles
  - e-CFv3.0 - European e-Competence framework for IT
  - CWA 16458 (2012): European ICT Professional Profiles Family Tree
  - ESCO (European Skills, Competences, Qualifications and Occupations) framework

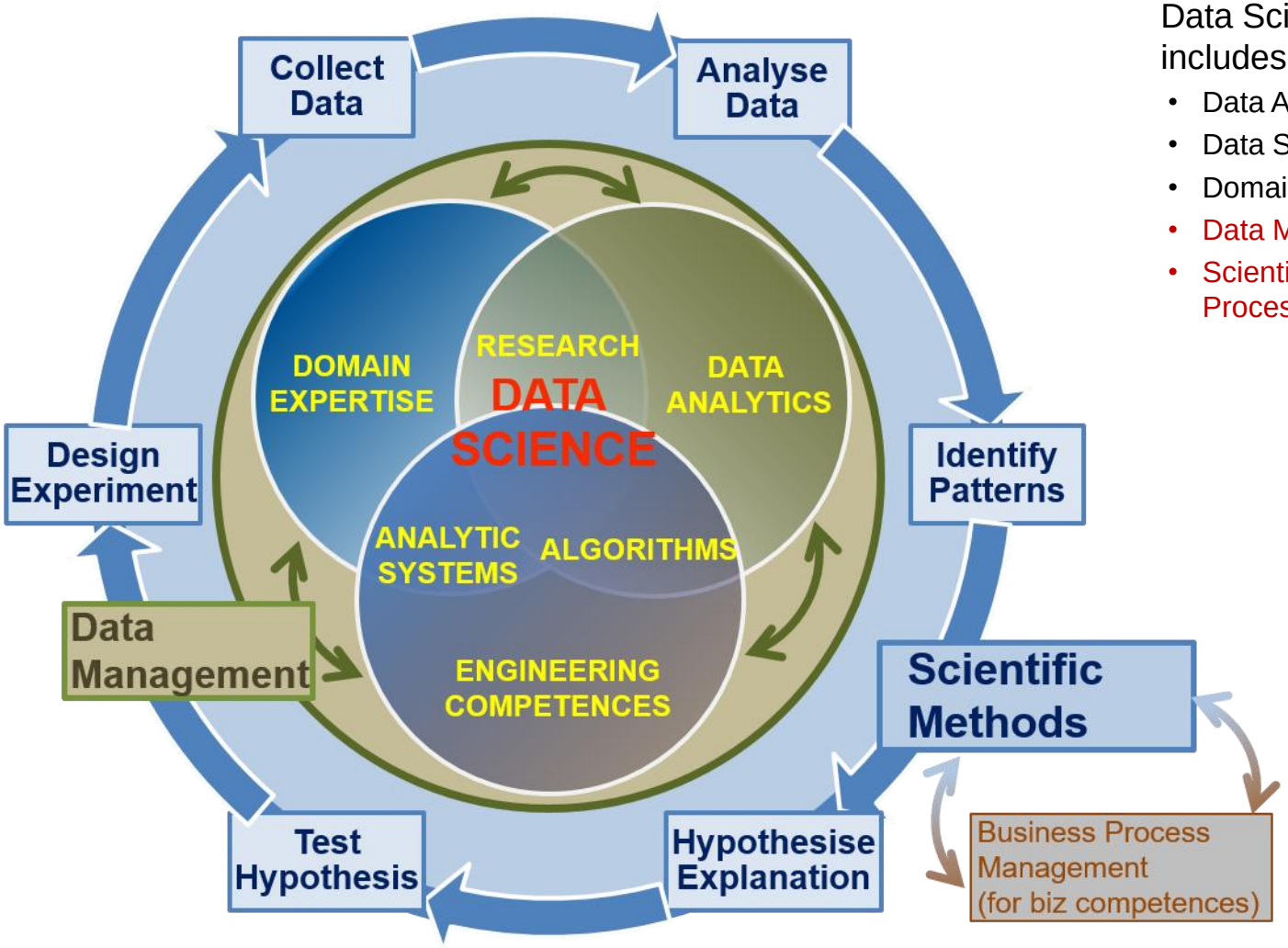
Identified Data Science Competence Groups



- Traditional/known Data Science competences/skills groups include
  - Data Analytics or Business Analytics or Machine Learning
  - Engineering or Programming
  - Subject/Scientific Domain Knowledge
- EDISON identified 2 additional competence groups demanded by organisations
  - **Data Management, Curation, Preservation**
  - **Scientific or Research Methods and/vs Business Processes/Operations**
- Other skills commonly recognized aka “soft skills” or “social intelligence”
  - Inter-personal skills or team work, cooperativeness
- All groups need to be represented in Data Science curriculum and training
  - Challenging task for Data Science education and training
- Another aspect of integrating Data Scientist into organisation structure
  - General Data Science (or Big Data) **literacy** for all involved roles and management
  - Common agreed and understandable way of communication and **information/data presentation**
  - ***Role of Data Scientist: Provide such literacy advice and guiding to organisation***

[ref] Legacy: NIST BDWG definition of Data Science

Data Science Competence Groups - Research



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

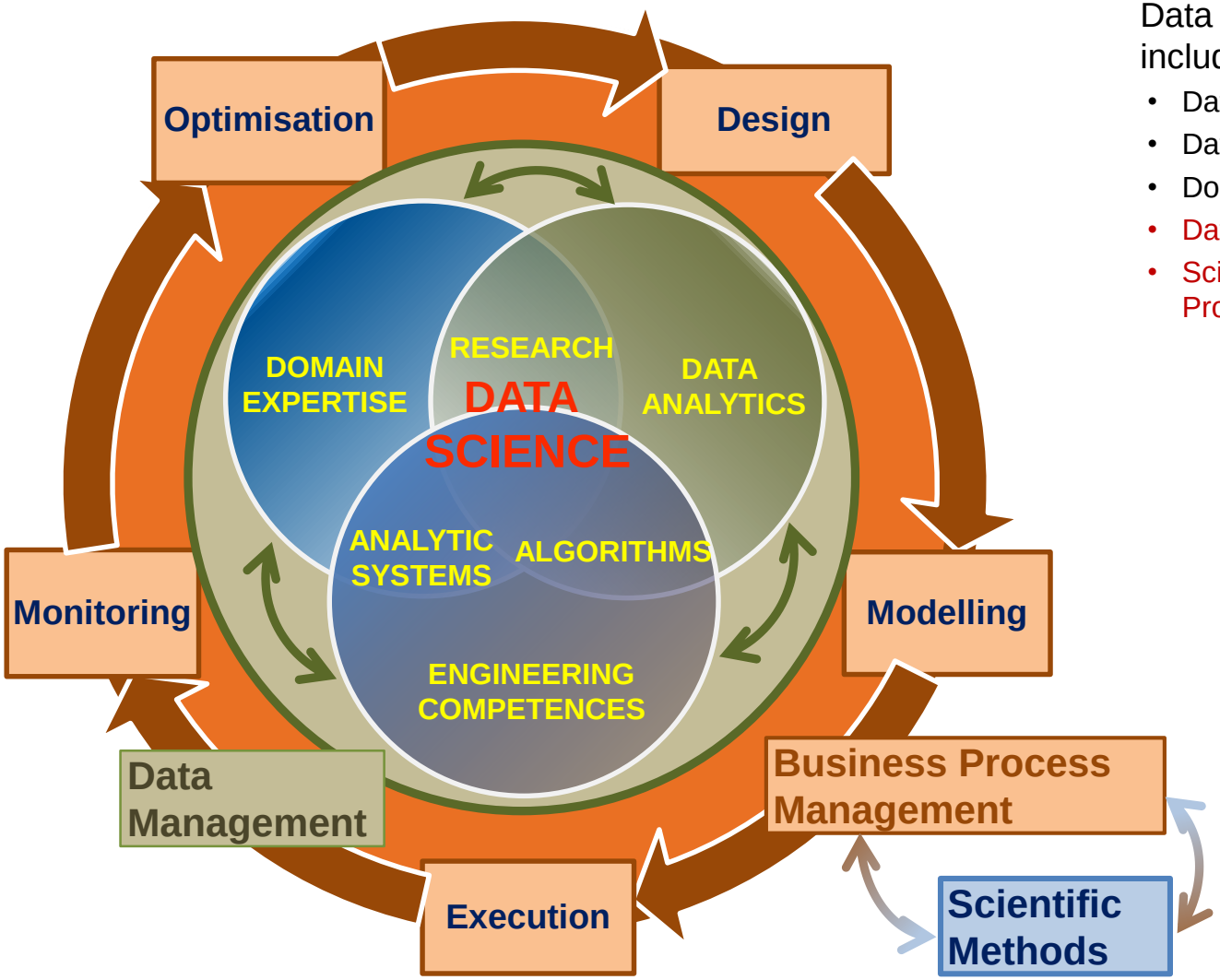
Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design

Data Science Competences Groups – Business



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Process Operations/Stages

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Identified Data Science Competence Groups

	Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
1	Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
3	Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
4	Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
5		**Collect and manage different source of data**	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
6		**Visualise complex and variable data.**	Develop algorithms to analyse multiple source of data	Influence the development of organizational objectives	Analyse multiple data sources for marketing purposes
7			Prototype new data analytics applications		Analyse customer data to identify/optimise customer relations actions

Identified Data Science Skills/Experience Groups

- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods
  - Application/subject domain related (research or business)
  - Mathematics and Statistics
- **Group 2: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Math & Stats apps & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - *Cloud based platforms and tools*
- **Group 3: Programming and programming languages and IDE**
  - General and specialized platforms for data analysis and statistics
- **Group 4: Soft skills or Social Intelligence**
  - Personal, inter-personal communication, team work (also called social intelligence or soft skills)

Identified Data Science Skill Groups

	Data Analytics and Machine Learning	Data Management/ Curation	Data Science Engineering (hardware and software)	Scientific/ Research Methods	Personal/Inter-personal communication, team work	Application/subject domain (research or business)
1	Artificial intelligence, machine learning	Manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources	Design efficient algorithms for accessing and analyzing large amounts of data	Interest in data science	Communication skills	Recommender or Ranking system
2	Machine Learning and Statistical Modelling	for data improvement	Big Data solutions and advanced data mining tools	Analytical, independent, critical, curious and focused on results	Inter-personal intra-team and external communication	Data Analytics for commercial purposes
3	Machine learning solutions and pattern recognition techniques	Data models and datatypes	Multi-core/distributed software, preferably in a Linux environment	Confident with large data sets and ability to identify appropriate tools and algorithms	Network of contacts in Big Data community	Data sources and techniques for business insight and customer focus
4	Supervised and unsupervised learning	Handling vast amounts of data	Databases, database systems, SQL and NoSQL	Flexible analytic approach to achieve results at varying levels of precision		Mechanism Design and/or Latent Dirichlet Allocation
5	Data mining	Experience of working with large data sets	Statistical analysis languages and tooling	Exceptional analytical skills		Game Theory
6	Markov Models, Conditional Random Fields	(non)relational and (un)-structured data	Cloud powered applications design			Copyright and IPR
7	Logistic Regression, Support Vector Machines	Cloud based data storage and data management				
8	Predictive analysis and statistics (including Kaggle platform)	Data management planning				
9	(Artificial) Neural Networks	Metadata annotation and management				
10	Statistics	Data citation, metadata, PID (*)				

Identified Big Data Tools and Programming Languages

	Big Data Analytics platforms	Math& Stats tools	Databases	Data/ applications visualization	Data Management and Curation platform
1	Big Data Analytics platforms	Advanced analytics tools (R, SPSS, Matlab, etc)	SQL and relational databases	Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)	Data modelling and related technologies (ETL, OLAP, OLTP, etc)
2	Big Data tools (Hadoop, Spark, etc)	Data Mining tools: RapidMiner, others	NoSQL Databases	Visualisation software (D3, Processing, Tableau, Gephi, etc)	Data warehouses platform and related tools
3	Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)	Mathlab	NoSQL, Mongo, Redis	Online visualization tools (Datawrapper, Google Charts, Flare, etc)	Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
4	Real time and streaming analytics systems (like Flume, Kafka, Storm)	Python	NoSQL, Teradata		Backup and storage management (iRODS, XArch, Nesstar, others
5	Hadoop Ecosystem/platform	R, Tableau R	Excel		
6	Spotfire	SAS			
7	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	Scripting language, e.g. Octave			
8	Amazon Data Analytics platform (Kinesis, EMR, etc)	Statistical tools and data mining techniques			
9	Other cloud based Data Analytics platforms (HortonWorks, Vertica LexisNexis HPCC System, etc)	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

- Big Data Analytics platforms
- Math& Stats tools
- Databases
- Data/applications visualization
- Data Management and Curation platform

Suggested e-CF extensions for Data Science

A. PLAN and Design
- A.10* Organisational workflow/processes model definition/formalisation
- A.11* Data models and data structures

B. BUILD: Develop and Deploy/Implement
- B.7* Apply data analytics methods (to organizational processes/data)
- B.8* Data analytics application development
- B.9* Data management applications and tools
- B.10* Data Science infrastructure deployment

C. RUN: Operate
- C.5* User/Usage data/statistics analysis
- C.6* Service delivery/quality data monitoring

15 Data Science Competences proposed covering different organizational roles and workflow stages
- Data Scientist roles are crossing multiple org roles and workflow stages

D. ENABLE: Use/Utilise
- D10. Information and Knowledge Management (powered by DS)
- D.13* Data presentation/visualisation, actionable data extraction
- D.14* Support business processes/roles with data and insight (support to D.5, D.6, D.7, D.12)
- D.15* Data management/preservation/curation with data and insight

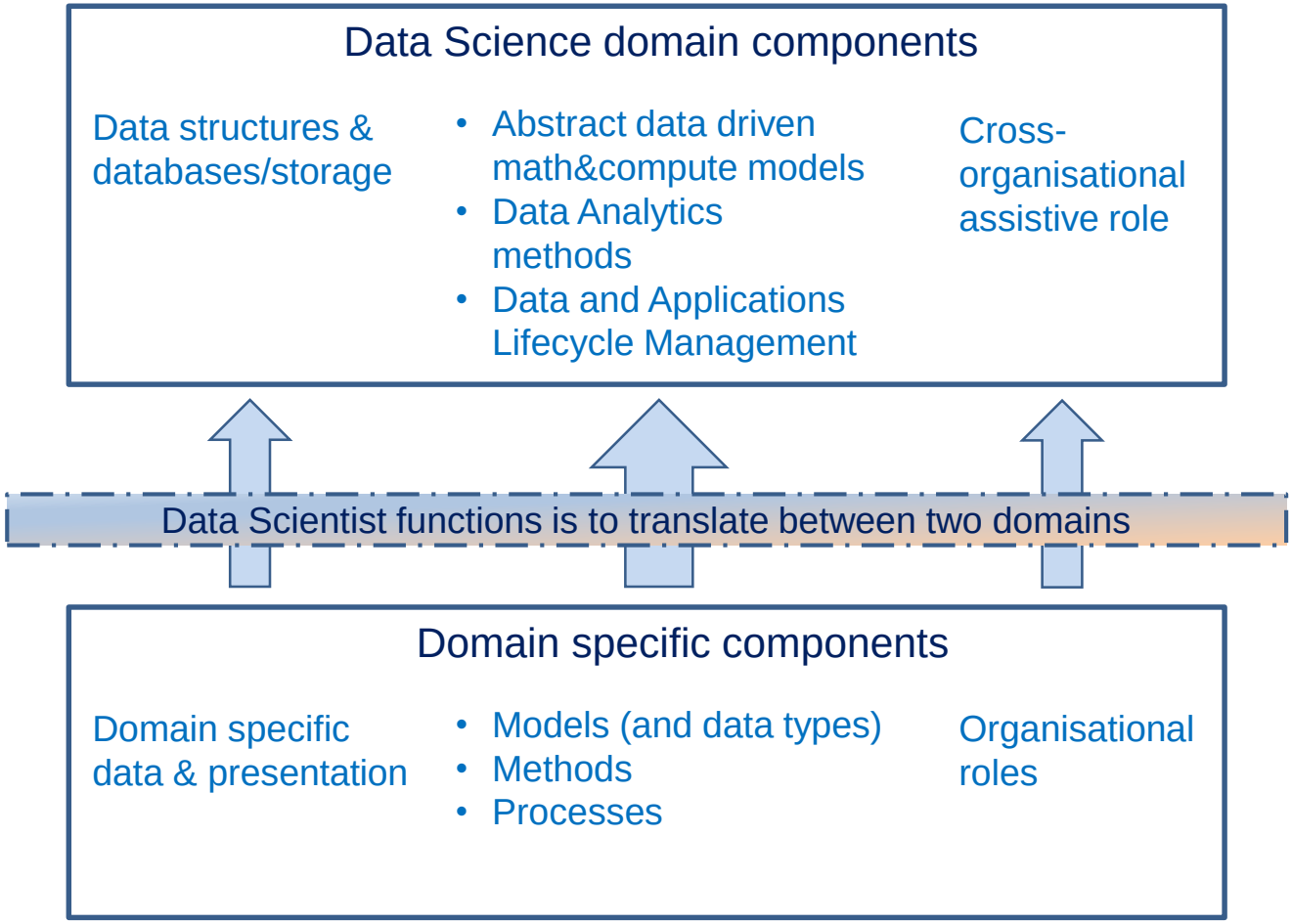
E. MANAGE
- E.10* Support Management and Business Improvement with data and insight (support to E.5, E.6)
- E.11* Data analytics for (business) Risk Analysis/Management (support to E.3)
- E.12* ICT and Information security monitoring and analysis (support to E.8)

Possible Data Scientist profiles/roles as extension to CWA16458 (2012) or ESCO

- Data Analyst, Business Analyst
  - Data Mining
  - Machine Learning
- Digital Librarian, Data Archivist, Data Curator, Data Steward
  - Data Management related competences
- Data Science Engineer/Administrator/Programmer
  - Data Analytics applications development
  - Scientific programmer
  - Data Science/Big Data Infrastructure engineer/developer/operator
- Data Science Researcher
  - Data Science creative
  - Data Science consultant/Analyst
- Data Scientist in subject/research domain

- Research e-Infrastructure brings its own specifics to required competences and skills definition

Data Science and Subject Domains



Data Scientist and Subject Domain Specialist

- **Subject domain components**
  - Model (and data types)
  - Methods
  - Processes
  - Domain specific data and presentation/visualization methods
  - Organisational roles and relations

- **Data Scientist is an assistant to Subject Domain Specialists**
  - Translate subject domain Model, Methods, Processes into abstract data driven form
  - Implement computational models in software, build required infrastructure and tools
  - Do (computational) analytic work and present it in a form understandable to subject domain
  - Discover new relations originated from data analysis and advice subject domain specialist
  - Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Data Science Body of Knowledge (DS-BoK)

DS-BoK Knowledge Area Groups (KAG)

- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Scientific/Research Methods group*
- KAG5-DSBP: Business process management group

- Data Science domain knowledge to be defined by related expert groups

KAG3-DSDM: *Data Management group: data curation, preservation and data infrastructure*

DM-BoK version2 “Guide for performing data management”
– 11 Knowledge Areas

(1) Data Governance,

(2) Data Architecture,

(3) Data Modelling and Design,

(4) Data Storage and Operations,

(5) Data Security,

(6) Data Integration and Interoperability,

(7) Documents and Content,

(8) Reference and Master Data,

(9) Data Warehousing and Business Intelligence,

(10) Metadata,

(11) Data Quality

Other Knowledge Areas motivated by RDA, European Open Data initiatives, European Open Data Cloud

(12) PID, ORCID

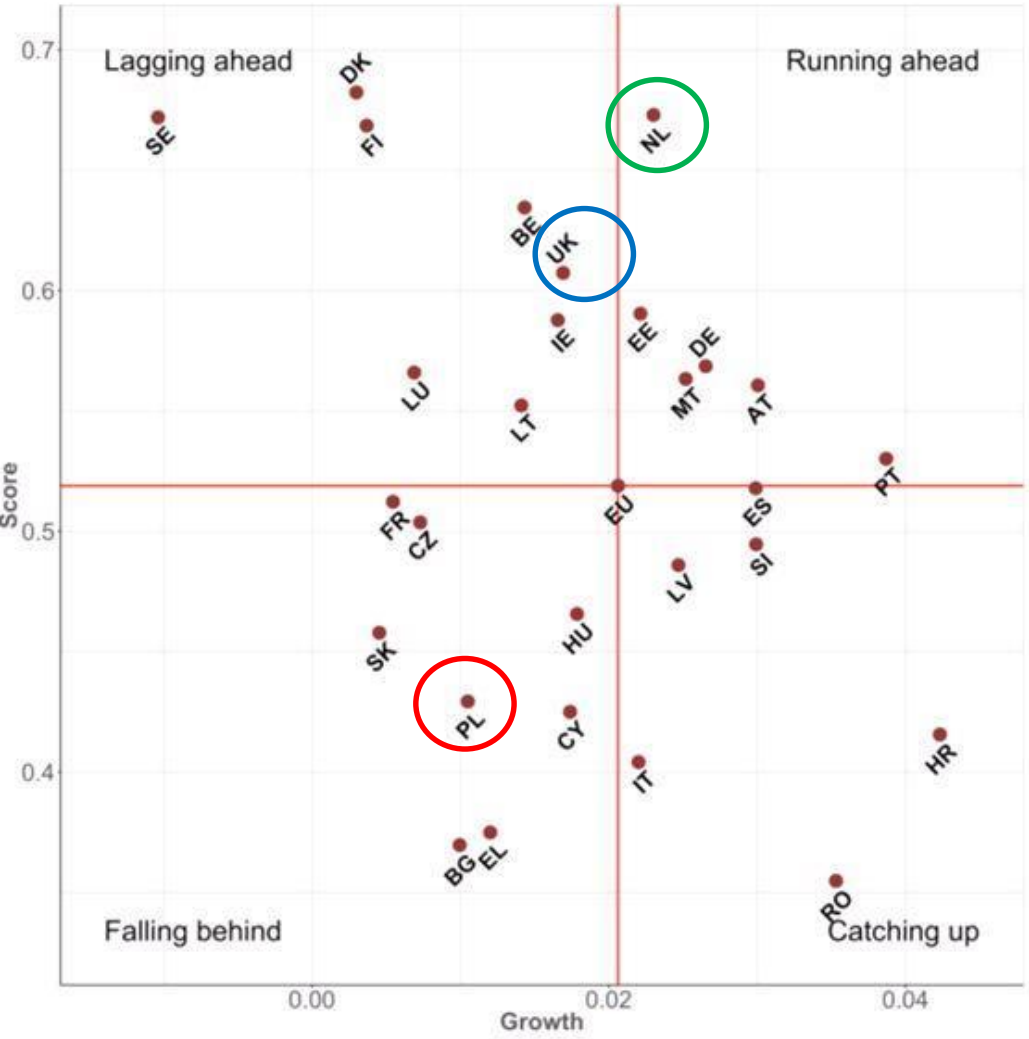
(13) Data Management Plan

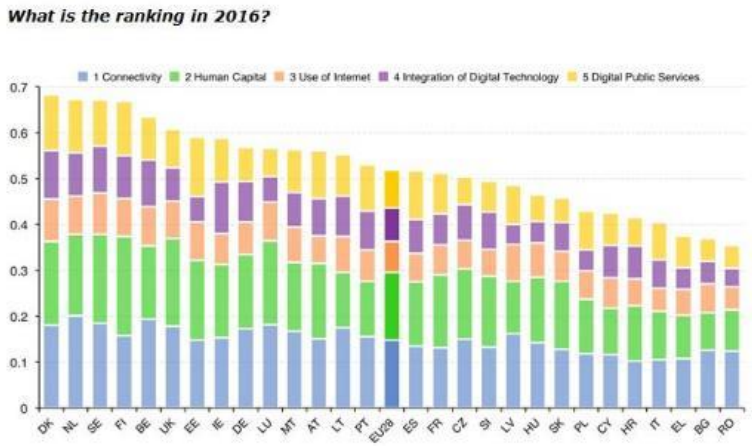
(14) Research Data Infrastructure

European Agenda on Skills for Digital Single Market (DSM) in Horizon 2020

- EC document to be published in May 2016 under Dutch presidency
- Multiple activities at EC
  - European Open Data Cloud (EODC) report by Barend Mons, Leiden University (bioinformatician)
    - To be published in April 2016
  - Standardisation for Big Data technologies Workshop 14 March 2016, Luxembourg
    - Call for more active contribution by European industry and experts in NIST Big Data WG and ISO/IEC JTC1 Big Data Study Group (SGBD)
  - eSkills workshop 16 March 2016, Den Haag
    - Addressing eSkills gap in Europe
  - Other events and activities by BDVA, OECD, etc

Digital Economy and Society Index (EU 2015-2016)





[ref] http://europa.eu/rapid/press-release_MEMO-16-385_en.htm

European consultation call – Deadline 30 April 2016

https://ec.europa.eu/futurium/en/content/consultation-european-e-infrastructure

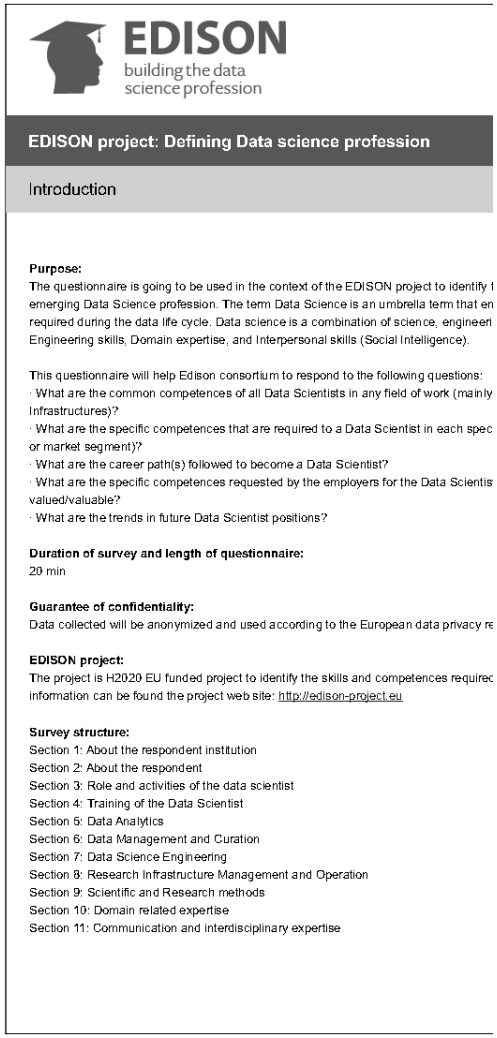
- What are the main challenges for the realisation of an integrated European e- infrastructure from the perspective of **scientific data-related needs** (from data access to sharing, analytics, re-use, preservation, standards, interoperability, value chain and other issues)?
- What are the challenges for reinforcing the **cooperation between European e-infrastructure service providers and their scientific users**, including thematic research infrastructures, to accelerate user's adoption of e-infrastructure services - such as identity management innovation - and foster innovation in e-infrastructures?
- What are the **challenges faced by industrial actors** preventing them to fully benefit from the services provided by European e-infrastructures and to contribute to the innovation of the existing e-infrastructures?
- **What are the main challenges Europe is facing regarding skills and competences required for effective data driven science, and management of research e-infrastructures?**

EDISON Project Engagement and Outreach

- EDISON Liaison Groups: Universities, Industry, Experts
- Champion universities
  - Summer 2016 workshop of Champions, Ambassadors and Adopters
- EDISON Survey on competences and skills for Data Science
  https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession
- Numerous workshops
- Data Science community portal http://www.edison-project.eu/
  - Community forum and community contribution
  - All major project deliverable are open for community discussion
  - Future: Personal profile building and competences self-assessment
- Future Data Science professional certification
  - For graduates and self-made data scientists

EDISON Survey: Data Science Competences and Skills

Survey link https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession



EDISON
building the data science profession

EDISON project: Defining Data science profession

Introduction

Purpose:
The questionnaire is going to be used in the context of the EDISON project to identify t
emerging Data Science profession. The term Data Science is an umbrella term that en
required during the data life cycle. Data science is a combination of science, engineeri
Engineering skills, Domain expertise, and Interpersonal skills (Social Intelligence).

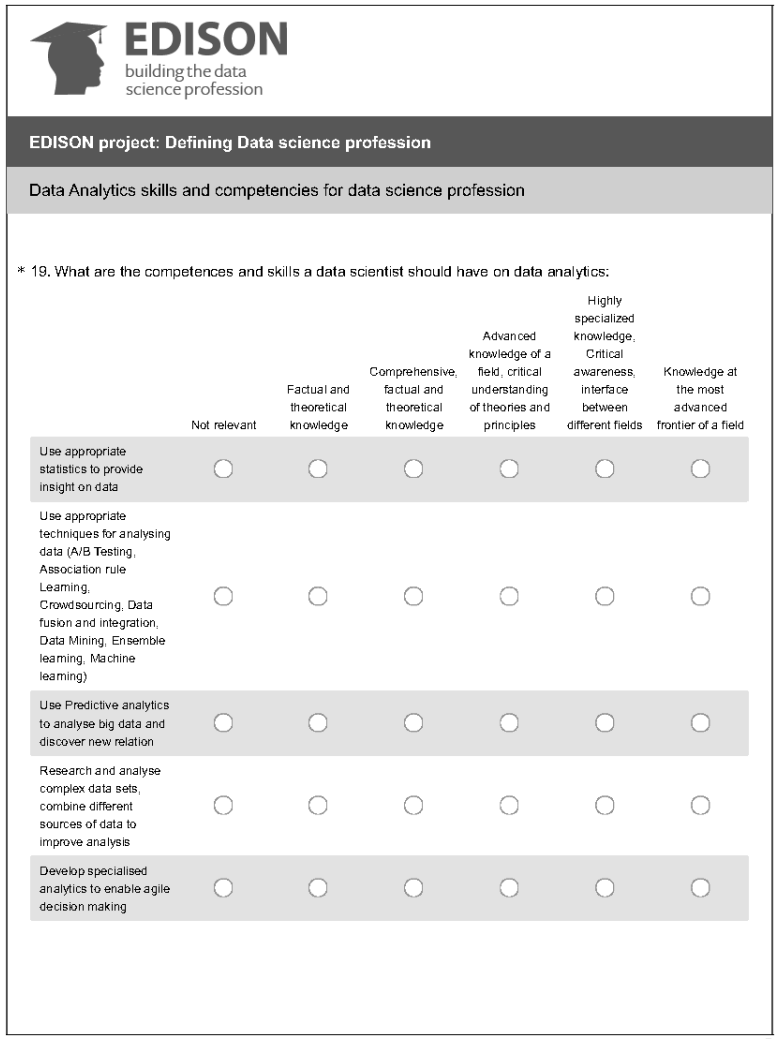
This questionnaire will help Edison consortium to respond to the following questions:
· What are the common competences of all Data Scientists in any field of work (mainly
Infrastructures)?
· What are the specific competences that are required to a Data Scientist in each spec
or market segment)?
· What are the career path(s) followed to become a Data Scientist?
· What are the specific competences requested by the employers for the Data Scientis
valued/valuable?
· What are the trends in future Data Scientist positions?

Duration of survey and length of questionnaire:
20 min

Guarantee of confidentiality:
Data collected will be anonymized and used according to the European data privacy re

EDISON project:
The project is H2020 EU funded project to identify the skills and competences requirec
information can be found the project web site: http://edison-project.eu

Survey structure:
Section 1: About the respondent institution
Section 2: About the respondent
Section 3: Role and activities of the data scientist
Section 4: Training of the Data Scientist
Section 5: Data Analytics
Section 6: Data Management and Curation
Section 7: Data Science Engineering
Section 8: Research Infrastructure Management and Operation
Section 9: Scientific and Research methods
Section 10: Domain related expertise
Section 11: Communication and interdisciplinary expertise



EDISON
building the data science profession

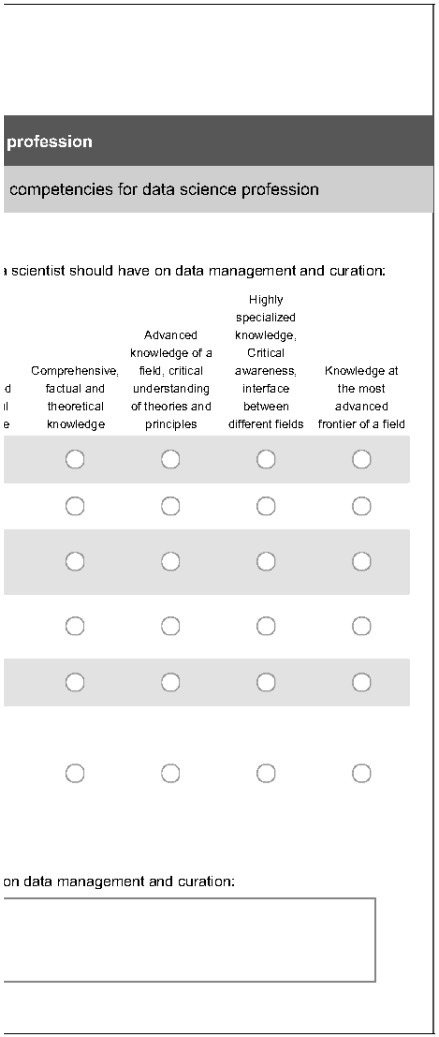
EDISON project: Defining Data science profession

Data Analytics skills and competencies for data science profession

* 19. What are the competences and skills a data scientist should have on data analytics:

	Not relevant	Factual and theoretical knowledge	Comprehensive, factual and theoretical knowledge	Advanced knowledge of a field, critical understanding of theories and principles	Highly specialized knowledge, Critical awareness, interface between different fields	Knowledge at the most advanced frontier of a field
Use appropriate statistics to provide insight on data	○	○	○	○	○	○
Use appropriate techniques for analysing data (A/B Testing, Association rule Learning, Crowdsourcing, Data fusion and integration, Data Mining, Ensemble learning, Machine learning)	○	○	○	○	○	○
Use Predictive analytics to analyse big data and discover new relation	○	○	○	○	○	○
Research and analyse complex data sets, combine different sources of data to improve analysis	○	○	○	○	○	○
Develop specialised analytics to enable agile decision making	○	○	○	○	○	○

8



profession

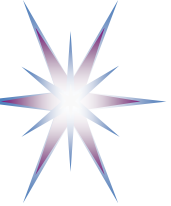
competencies for data science profession

scientist should have on data management and curation:

	Comprehensive, factual and theoretical knowledge	Advanced knowledge of a field, critical understanding of theories and principles	Highly specialized knowledge, Critical awareness, interface between different fields	Knowledge at the most advanced frontier of a field
	○	○	○	○
	○	○	○	○
	○	○	○	○
	○	○	○	○
	○	○	○	○
	○	○	○	○

on data management and curation:

10



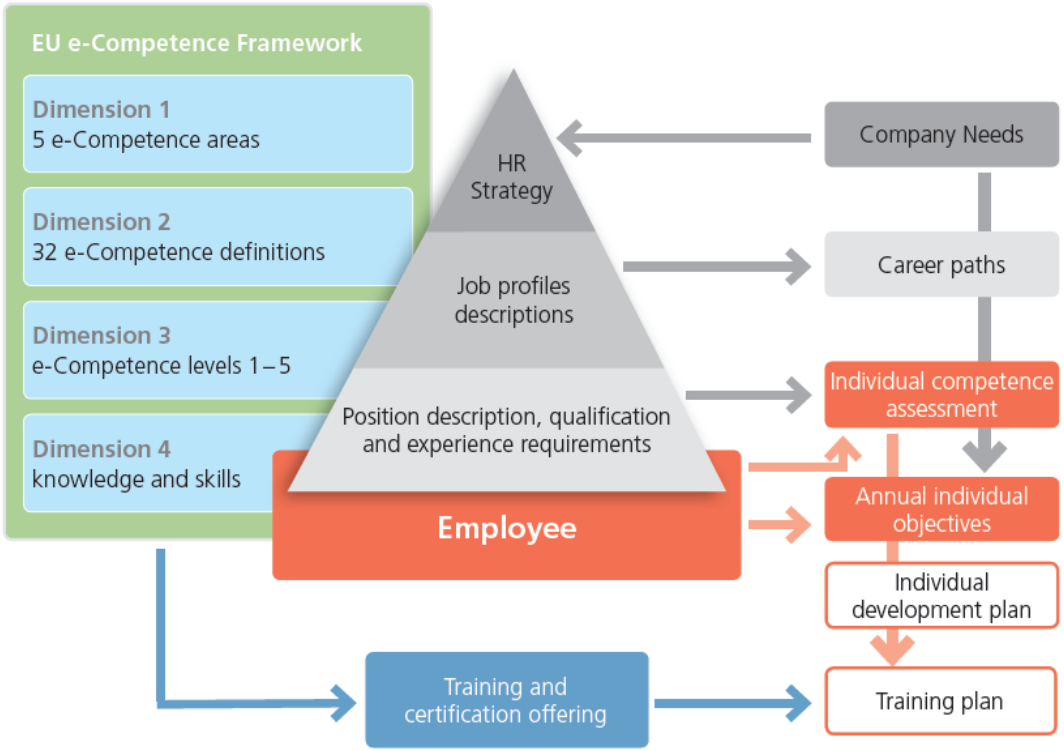
Questions and Discussion

Additional information

- EDISON Approach: e-CFv3.0 and CF-DS
- Data Science occupations in ESCO taxonomy

EDISON Approach: e-CFv3.0 and CF-DS

- Competence Framework for Data Science (CF-DS) definition will be built based on European e-Competence framework for IT (e-CFv3.0)
  - Linking *scientific research cycle/flow*, organizational roles, competences, skills and knowledge
  - Defining *Data Science Body of Knowledge (DS-BoK)*
  - Mapping CF-DS and DS-BoK to academic disciplines in a DS *Model Curriculum (MC-DS)*



- Multiple use of e-CFv3.0 within ICT organisations
- Provides basis for individual career path, competence assessment, training and certification

- EDISON CF-DS will be used for defining DS-BoK and MC-DS, linking organizational functions and required knowledge
- Provide basis for individual (self) training and certification

e-CFv3.0 Internal Structure: Refactoring for CF-DS

European e-Competence Framework 3.0 overview

Dimension 1 5 e-CF areas (A – E)	Dimension 2 40 e-Competences identified	Dimension 3 e-Competence proficiency levels e-1 to e-5, related to EQF levels 3–8				
		e-1	e-2	e-3	e-4	e-5
A. PLAN	A.1. IS and Business Strategy Alignment				■	■
	A.2. Service Level Management			■	■	
	A.3. Business Plan Development			■	■	■
	A.4. Product/Service Planning		■	■	■	
	A.5. Architecture Design			■	■	■
	A.6. Application Design	■	■	■		
	A.7. Technology Trend Monitoring				■	■
	A.8. Sustainable Development			■	■	
	A.9. Innovating				■	■
B. BUILD	B.1. Application Development	■	■	■		
	B.2. Component Integration		■	■	■	
	B.3. Testing	■	■	■	■	
	B.4. Solution Deployment	■	■	■		
	B.5. Documentation Production	■	■	■		
	B.6. Systems Engineering			■	■	
C. RUN	C.1. User Support	■	■	■		
	C.2. Change Support		■	■		
	C.3. Service Delivery	■	■	■		
	C.4. Problem Management		■	■	■	
D. ENABLE	D.1. Information Security Strategy Development				■	■
	D.2. ICT Quality Strategy Development				■	■
	D.3. Education and Training Provision		■	■		
	D.4. Purchasing		■	■	■	
	D.5. Sales Proposal Development		■	■		
	D.6. Channel Management			■	■	
	D.7. Sales Management			■	■	■
	D.8. Contract Management		■	■	■	
	D.9. Personnel Development		■	■	■	
	D.10. Information and Knowledge Management			■	■	■
	D.11. Needs Identification			■	■	■
	D.12. Digital Marketing		■	■	■	
E. MANAGE	E.1. Forecast Development			■	■	
	E.2. Project and Portfolio Management		■	■	■	■

- 4 Dimensions
  - Competence Areas
  - Competences
  - Proficiency levels
  - Skills and Knowledge
- 5 Competence Area defined by ICT Business Process stages
  - Plan
  - Build
  - Run
  - Enable
  - Manage

-> Refactor to Scientific Research cycle/workflow (and linked to Scientific Data Lifecycle)

  - See example of RI manager at IG-ETRD wiki and meeting

- Each competence has 5 proficiency level
  - Ranging from technical to engineering to management to strategist/expert level
- Knowledge and skills property are defined for/by each competence and proficiency level (not unique)

Definitions (according to e-CFv3.0)

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
  - Competence vs Competency (e-CF vs ACM)
    - Competence is ability acquired by training or education (linked to learning outcome)
    - Competency is similar to skills or experience (acquired feature of a person)
  - Competence can be treated as outcome of learning or training

- **Knowledge** in the context of competence definition is treated as something to know, to be aware of, familiar with, and obtained as a part of education.

- **Skills** is treated as provable ability to do something and relies on the person's experience.

Data Science occupations in ESCO taxonomy (1)

Professionals				
	Science and engineering professionals	Data Science Professionals	Data Science professionals not elsewhere classified	Data Scientist
				Data Science Researcher
				(Big) Data Analyst
				Data Science (Application) Programmer
				Business Analyst
		Database and network professionals	Large scale (cloud) data storage designers and administrators	Large scale (cloud) database designer*)
			Database designers and administrators	Large scale (cloud) database administrator*)
			Database and network professionals not elsewhere classified	Scientific database administrator*)
	Information and communications technology professionals	Data Science technology professionals	Data handling professionals not elsewhere classified	Digital Librarian
				Data Archivist
				Data Steward
				Data curator

Data Science occupations in ESCO taxonomy (2)

Technicians and associate professionals				
	Science and engineering associate professionals	Data Science Technology Professionals	Data Infrastructure engineers and technicians	Big Data facilities Operators
				Large scale (cloud) data storage operators
			Database and network professionals not elsewhere classified	Scientific database operator*)
Managers				
	Production and specialised services managers	Data Science/Big Data Infrastructure Managers		Data Science/Big Data Infrastructure Manager
			Research Infrastructure Managers	RI Manager
				RI Data storage facilities manager
Clerical support workers				
	General and keyboard clerks			
	Data handling support workers (alternative)	Data and information entry and access	Digital Archivists and Librarians	Digital Librarian
				Data Archivist
				Data Steward
				Data curator