Big Data Platforms and New Profession of Data Scientist

Yuri Demchenko
System and Network Engineering Group
University of Amsterdam
Amsterdam, The Netherlands
e-mail: y.demchenko@uva.nl

***Abstract*—Big Data are becoming a new technology focus both in science and in industry and motivate technology shift to data centric architecture and operational models. There is a vital need to define the basic information/semantic models, architecture components and operational models that together comprise a so-called Big Data Ecosystem. This paper discusses a nature of Big Data and proposes the Big Data Architecture Framework (BDAF) definition that includes the following components: Big Data Infrastructure, Big Data Analytics, Data structures and models, Big Data Lifecycle Management, Big Data Security. The paper analyses requirements BDAF components and provides overview of cloud based Big Data platforms. The paper also addresses demand for a new profession of Data Scientist and describes identified competences and skills for Data Scientists. The paper refers to the EDISON project that develops a number of components to create a foundation for establishing a new profession in Europe.**

Keywords- Big Data Architecture Framework (BDAF), Cloud based Big Data platforms, Data Scientist Profession, Data Science Competence Profile.

I. INTRODUCTION

Big Data, also referred to as Data Intensive Technologies, are becoming a new technology trend in science, industry and business [1, 2, 3]. Big Data are becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer. Current technologies such as Cloud Computing and ubiquitous network connectivity provide a platform for automation of all processes in data collection, storing, processing and visualization.

Big Data technologies developments and increasing adoption of data driven technologies by science and industry motivates creation of the new profession of Data Scientist who is expected to drive use of Big Data by companies and extract actionable value from data that are increasingly collected by companies and available from multiple sources of information both publicly available and provided as different data services. The Data Science profession establishing for Europe with international recognition is a subject of the H2020 funded project EDISON [4].

II. BIG DATA DEFINITION AND ANALYSIS

To reflect the new properties of the Big Data as a driving force for new data driven technologies in research and industry we proposed an extended Big Data definition that includes five parts [5]:

(1) Big Data Properties: 6V
- Volume, Variety, Velocity, Value, Veracity, Variability

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

(3) New Analytics
- Highly scalable, real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
- Cloud based infrastructure, storage, network, high performance computing
- New data centric service and security models

(5) Source and Target that are important aspect defining domain related data types and data models
- Fully digitised input and output, (ubiquitous) sensor networks, full digital control

III. BIG DATA ARCHITECTURE FRAMEWORK

In our previous works we proposed the Big Data Architecture Framework (BDAF) that support the extended Big Data definition and support the main components and processes in the Big Data (eco)systems that is compliant with the NIST Big Data Interoperability Framework [3]. The proposed BDAF comprises of the following 5 components:

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

(2) Big Data Management
- Big Data Lifecycle (Management), provenance
- Data linkage, Curation, Archiving

(3) Big Data Analytics and Tools
- Big Data analytics and applications

(4) Big Data Infrastructure (BDI)
- Big Data Storage, Compute/HPC, Network
- Big Data analytics infrastructure and platforms

(5) Big Data Security
- Security data storage, transfer, trusted processing environment

IV. BIG DATA INFRASTRUCTURE (BDI) AND PLATFORMS

Figure 1 provides a general view on the Big Data infrastructure that includes the general infrastructure for general data management, typically cloud based, and Big Data Analytics infrastructure that will require specialized and high-performance computing clusters.

General BDI services and components include
- Big Data Management tools
- Registries, indexing/search, semantics, namespaces
- Security infrastructure (access control, policy enforcement, confidentiality, trust, availability, privacy)
- Collaborative environment (groups management)

We define Federated Access and Delivery Infrastructure (FADI) as an important component of the general BDI that interconnects different components of the cloud/Intercloud based infrastructure combining dedicated network connectivity provisioning and federated access control

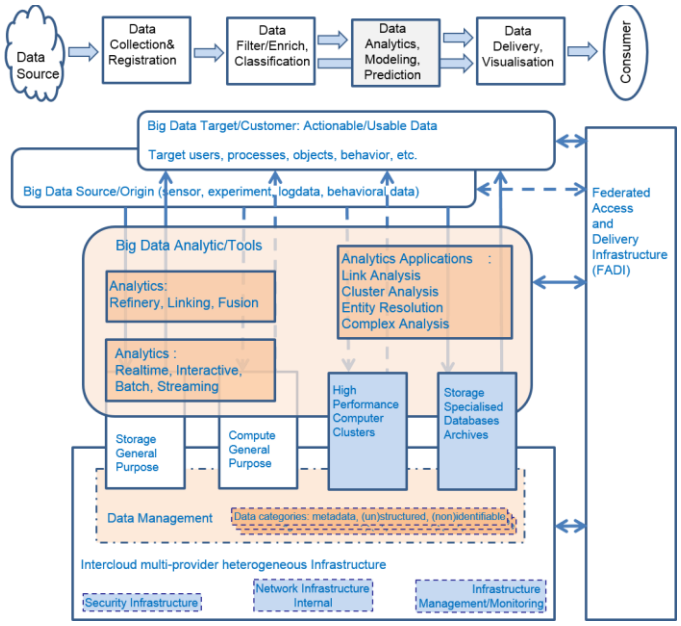

Figure 1. General Big Data Infrastructure functional components

Besides the general cloud base infrastructure services (storage, compute, infrastructure/VM management) the following specific applications and services are required to support Big Data and other data centric applications:

- Hadoop based services and tools, streaming analytics, etc.
- Specialist data analytics tools (events, data mining, etc.)
- Databases/Servers SQL, NoSQL

Big Data analytics tools are currently offered by the major cloud services providers such as: Amazon Elastic MapReduce and Dynamo, Microsoft Azure HDInsight, IBM Big Data Analytics. Scalable Hadoop and data analytics tools services are offered by few companies that position themselves as Big Data companies such as Hortonworks, Cloudera, and others.

V. Data Science Competence Framework and Body of Knowledge

The EU funded H2020 EDISON project [4] targets to formally define the Data Science profession in terms of defining the Data Science Competences Framework (CF-DS), Body of Knowledge (DS-BoK), Model Curriculum (MC-DS), and professional certification scheme.

The EDISON CF-DS development follows the European e-Competences Framework (e-CF3.0) guiding principles [6]. The initial study has identified the following groups of competences (see Figure 2):

- Data Analytics (including Machine Learning and Business Analytics specifically for business applications)
- Data Science Engineering (including both Data Science Infrastructure and Data Science applications engineering)
- Subject/Scientific Domain Knowledge
- Data Management, Curation, Preservation - *new*
- Scientific or Research Methods (for science) and Business Process Management (for business and industry) - *new*

where the newly identified competence areas provide a better basis for defining education and training programmes for Data Science related jobs, re-skilling and professional certification. Data Management, curation and preservation are important components of European Research Area policy.

Knowledge of the scientific research methods and techniques makes the Data Scientist profession different from all previous professions. Scientific research typically includes collection of data in passive observation or active experiments which aim to verify one or another scientific hypothesis. Scientific research and discovery methods are typically based on the initial hypothesis and a model which can be refined based on the collected data. The refined model may lead to a new more advanced and precise experiment and/or the previous data re-evaluation.

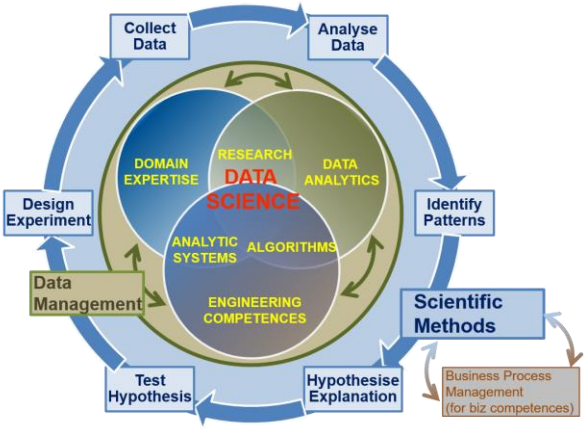

Figure 2. Data Science competences groups.

The CF-DS competences are mapped to required knowledge areas and define the learning outcome in a curriculum providing a basis for the Data Science Body of Knowledge (DS-BoK) and curriculum definition [7], providing also a basis for knowledge assessment and professional certification.

References

[1] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[2] 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*. [Online]. Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] NIST SP 1500-1 NIST Big Data Interoperability Framework (NBDIF): Volume 1: Definitions, Sept 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[4] EDISON Project: Building Data Science Profession [online] http://www.edison-project.eu/

[5] Demchenko, Yuri, Peter Membrey, Cees de Laat, Defining Architecture Components of the Big Data Ecosystem. Second International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2014). Part of The 2014 Int. Conf. on Collaboration Technologies and Systems (CTS 2014), May 19-23, 2014, Minneapolis, USA.

[6] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-

[7] Demchenko, Y., E.Gruengard, S.Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. In Proc. 6th IEEE Intern Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 Dec 2014, Singapore