Request for Change

Date: 13 April 2021

Requestor: NATO Communications and Information Agency

POC: Michael STREET

Email: michael.street@ncia.nato.int

Target Version: SFIA v8

Rational:

SFIA v7 identifies a range of roles which collectively provide the 'data science' function. While it is true that no one individual can adequately deliver 'data science', it is also the case that many organisations hire data scientists, and many individuals describe themselves thus.

In relation to previously existing professions, the data science field can be thought of as a superset of business intelligence, data mining, data analytics and that includes elements from software engineering and data engineering.

The UK Government's Digital Service (GDS) defines the data scientist's role as a member of a multidisciplinary team, working with data architects, data engineers, data analysts and others. They state that, "A Data Scientist uses data to identify and solve complex business problems. They have an interdisciplinary focus, using techniques and knowledge from a range of scientific and computer science disciplines (for example, statistics, analytics, machine learning). A Data scientist is open and transparent, collaborating with others to share good practice and continuously improve outputs."

Data science jobs typically involve developing machine learning, statistical, deep learning and natural language processing (NLP) models in languages like Python, R, SQL, Scala and Java, using frameworks such as Scikit-Learn, Spark ML, Tensorflow and PyTorch.

The three most commonly used data science job titles on LinkedIn are, in ascending order of seniority: "Data Scientist" (grade-1), "Senior Data Scientist" (grade-2), and "Principal Data Scientist" (grade-3). However, there is no consensus on the number of different types of data scientist across industry or academia, although recurring categories include:

- *Generalist: Applies machine learning and statistics to develop models that generate insight to solve business problems. Generalist roles tend to be advertised as grade-1 or grade-2 positions, without any requirement for specialist skills or domain-specific knowledge.*
- *Specialist: Extends beyond the capabilities of a Generalist, focusing on specialist machine learning skills (e.g. deep learning, natural language processing, audio/image processing) and application domains (e.g. healthcare, finance, manufacturing). Specialist roles tend to be advertised as grade-2 or grade-3 positions.*
- *Researcher: Focuses on the development of novel machine learning algorithms or application of existing algorithms in new settings. These tend to be advertised as grade-2 or grade-3 positions.*

There are a several certifications that are specific to data science, many of which are vendor-specific:

- *Data Science Council of America (DASCA) Senior Data Scientist (SDS)*
- *Data Science Council of America (DASCA) Principal Data Scientist (PDS)*
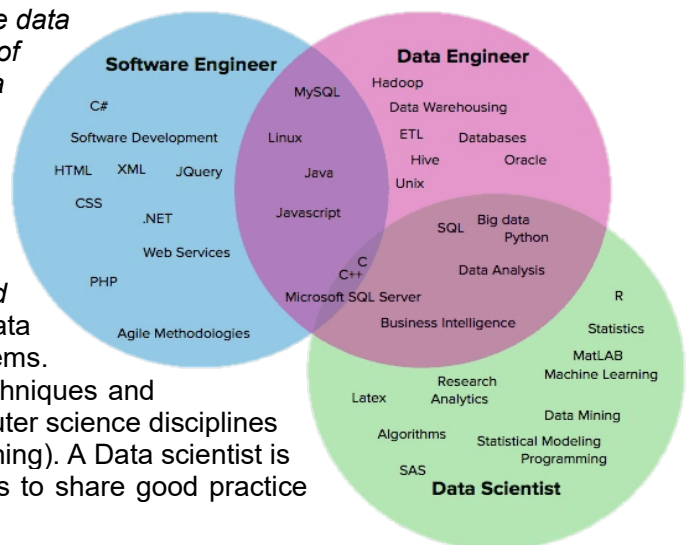
- *Dell EMC Data Science Track (EMCDS)*
- *IBM Data Science Professional Certificate*
- *Microsoft Certified: Azure Data Scientist Associate*
- *Open Certified Data Scientist (Open CDS)*
- *SAS Certified Data Scientist*

*It is suggested to add a new skill "Data Scientist" at levels 4,5 and 6 to the SFIA Framework, corresponding to the commonly used grade-1, grade-2 and grade-3 respectively. The following proposed competency descriptions are an extension of the SFIA **Analytics** INAN to include additional competencies commonly associated with these roles in NATO and also outlined in the SFIA discussion on data science. Our emphasis is on incorporating competencies from Innovation, Data Management, Programming/software development, and Data Visualisation. Additions or changes to the original Analytics INAN text are highlighted in green.*

Data Science (DATS)

Overall description:

The application of mathematics, statistics, predictive modelling and machine learning techniques to discover meaningful patterns and knowledge in recorded data. The planning, designing, creation, amending, verification, testing and documentation of new and amended software components in order to deliver agreed value to stakeholders. Analysis of data with high volumes, velocities and variety (numbers, symbols, text, sound and image). Development of forward-looking, predictive, real-time, model-based insights to create value and drive or automate effective decision-making. Presenting findings and data insights in creative ways to facilitate the understanding of data across a range of technical and non-technical audiences. The identification, validation and exploitation of internal and external data sets generated from a diverse range of processes. The management of data and information in all its forms and the analysis of information structure (including logical analysis of taxonomies, data and metadata). Adherence to legal requirements and ethical guidelines for data privacy and automated decision systems. The capability to identify, prioritise, incubate and exploit innovation opportunities provided by data science technology.

Level 4 description:

Applies a range of mathematical, statistical, predictive modelling or machine learning techniques in consultation with experts if appropriate, and with sensitivity to the limitations of the techniques. Designs, codes, verifies, tests, documents, amends and refactors complex programs/scripts and integration software services. Selects, acquires and integrates data for analysis. Develops data hypotheses and methods, trains and evaluates analytics models, shares insights and findings and continues to iterate with additional data. Applies a variety of visualisation techniques and designs the content and appearance of data visuals. Follows legal requirements and ethical guidelines for data privacy and automated decision systems. Takes responsibility for the accessibility, retrievability, security, quality, retention and ethical handling of specific subsets of data.

Level 5 description:

Evaluates the need for analytics, assesses the problems to be solved and what internal or external data sources to use or acquire. Specifies and applies appropriate mathematical, statistical, predictive modelling or machine learning techniques to analyse data, generate insights, create value and support decision-making. Establishes the purpose and parameters of the data visualisation. Advises on and provides overall control, to ensure appropriate use of data visualisation tools and techniques. Takes technical responsibility across all stages and iterations of analytics software development. Measures and monitors applications of project/team standards for analytics software construction including software security. Manages reviews of the benefits and value of analytics techniques and tools and recommends improvements. Manages the data

science innovation pipeline and executes innovation in analytics processes. Contributes to the development of data science policy, standards and guidelines. Communicates and follows legal requirements and ethical guidelines for data privacy and automated decision systems. Devises and implements ~~master~~ data management processes, including classification, security, quality, ethical principles, retrieval and retention processes.

Level 6 description:

Develops data science policy, standards and guidelines. Establishes and manages analytics methods, techniques and capabilities to enable the organisation to analyse data, to generate insights, create value and drive decision-making. Plans and leads software construction activities for strategic, large and complex analytics development projects. Obtains organisational commitment to data analytics and data science innovation. Sets direction and leads the introduction, innovation and use of analytics to meet overall business requirements, ensuring consistency across all user groups. Identifies and establishes the veracity of the internal and external sources of information which are relevant to the operational needs of the enterprise. Ensures compliance with legal requirements and ethical guidelines for data privacy and automated decision systems. Derives an overall strategy of ~~master~~ data management, within an established information architecture, that supports the development and secure operation of data science services.