**Unit 5 Reading - Data Cleaning and Automating Data Collections**

**IBM Intellas. (2016) Qradar and KAIF Integration Report.**  
<https://www.intellas.biz/case-studies/ibm-qradar-kaif-integration>

The article discusses a tool which uses a range of algorithms, methods and techniques that can facilitate the identification of useful data from the huge amount collected as part of ICT infostructure to support cyber security intelligence and evidence gathering plus analysis.

**Datanami. (2019) Data Pipeline Automation: The Next Step Forward in DataOps.**  
<https://www.datanami.com/2019/04/24/data-pipeline-automation-the-next-step-forward-in-dataops/>

Data pipeline means all the collecting, transforming and preparing data for advanced analytics and machine learning. Automating that workload is the next step forward

Data engineers some of the most in demand people in organisations, Data Ops tools allow them to manage the data pipeline. Whilst this data pipeline remains manual and reliant on individuals it is not responsive to the organisations needs in a data driven business.

Emerging DataOps borrows from DevOps including focus on agility, leanness and continuous delivery. Key difference that within a data analytics environment with many tools, sources and warehouses.

As complexity increases and development team size grows, organisations need processes to govern data lifecycle from ingestion to analysis. Aim to increase agility and cycle time and improve data quality and output.

New Dataops platforms aim to accelerate the work of data engineers and make them more productive. They make data pipeline work re-usable and help with repetitive tasks. Automation of complex work is possible. Examples include data transformation tasks, managing schemas, automating the creation of a SQL Server data warehouse. Monitoring data pipeline performance and data integrity. Pre-built connectors to common data sources.

**Reflection:**

Management of the data pipeline is hugely important in data analytics projects and data science. Being able to manage and automate those processes via Data Ops can accelerate that process and make it more re-usable and resilient.