NEXT GENERATION IBM DATASTAGE - ON CLOUD, AGILE AND AFFORDABLE

Presented by

IBM

Smallnet
Integration • Governance • Warehousing

01	Introductions
02	DataStage today, and marketplace drivers
03	The NEXT Generation of DataStage – NEW developments, NEW User Interface with an interactive DEMO of the software.
04	Flexible deployment and licensing options – from fully managed by IBM to traditional deployment options with hybrid and dual entitlements
05	Questions and Answer Session

Presenters

PETER KEITH

IBM Brand Technical Specialist Data & AI

GEMMA WOODCOCK

IBM Senior Data & AI Technical sales

PAUL RANSON

SmallNet Consulting

Logistics & Disclaimer

The recording will be available approximately 24 hours after the webinar. Additionally, the slides being presented (has access to various assets) will be available in PDF format and these they can be accessed after the webinar via Barry O'Mahoney at Barry.omahoney@ie.ibm.com or Paul Ranson at paul.ranson@smallnetconsulting.co.uk

This DataStage Webinar is being organised and run in-conjunction with our accredited Business partner Smallnet consulting and as part of the follow up we would share the delegates attending or registered with them, but should there be any issues with this please advise IBM (Barry.omahoney@ie.ibm.com)

Smallnet Overview

21 Years of Information Management Solutions

Governance Integration Warehousing Quality

IBM Business Partner Trusted Supplier Cross Industry

Software | Services | Training | Solutions

> We have been working with Data Stage &Information Server portfolio longer than IBM have owned the software

SmallNet website; https://www.smallnetconsulting.co.uk/

Case Studies and Blogs; https://www.smallnetconsulting.co.uk/resources/

DataStage Today and the Market Place

- **TWENTY ONE Years plus** - Built on 21 plus years of experienced developers and market trends adding value and functionality, plus integration with many modules ie Quality, Catalog, Replication
- **10,000 PLUS USERS** - Today has over 10,000 users in many, many mission critical applications ie Finance, Insurance, Retail, CPG and many more industries
- **PARALLEL ENGINE** - DataStage is the only ETL/ELT product in the Gartner Magic Quadrant that was designed and built with a 'parallel engine' built in as standard.
- **DESIGN ONCE USE ANYWHERE** - DataStage can and has been all through the technology iterations of Centralised, De-Centralised, GRID, Hadoop, Data Lakes and Cloud with the same product code
- **RESOURCES** - DataStage (Information Server) has 1,000 of skilled professionals in the market place so you can rely on a strong eco system for training, development and deployment.
- **MARKET DRIVERS: DATA FABRIC/MESH/CLOUD/OBSERVABILITY** - IBM Watson Knowledge Catalog brings in Data Quality to merge with the portfolio. Plus NEW DATABAND Offering….

POLLING QUESTION 1

Which best describes your use of DataStage – Current and Planned

- Maintaining only long-standing solution, no growth
- Continuous new development and growth of DataStage use
- Considering using DataStage
- Sunsetting DataStage use, looking to use alternative solutions

The NEXT Generation of DataStage - NEW developments, NEW User Interface with an interactive DEMO of the software.

PETER KEITH – IBM Brand Technical Specialist Data & AI

Next Generation DataStage

Modernizing the industry's most performant data integration solution

Modernized, cloud-native ETL tool with 20+ years experience

- DataStage, the industry leader for data integration, was re-architected and is available as a true cloud-native, cloud-first experience for our customers.

Different data integration styles with polyglot execution engines

- Combine batch integration with other capabilities in Cloud Pak for Data to actualize data integration styles like parallel processing, virtualization, replication, streaming, and preparation.

Multicloud scalability, elasticity, and runtime execution

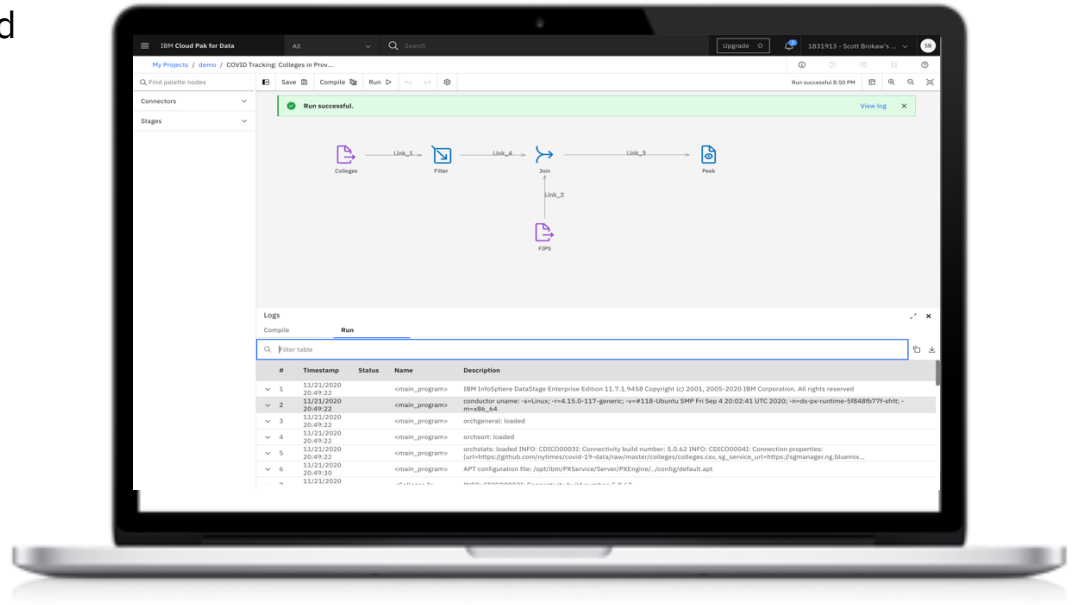
- Design once, dynamically run anywhere with built-in automatic workload balancing, parallelism and dynamic scalability.
- Increase data gravity with distributed processing for hybrid-cloud or multicloud requirements.

Integrated with IBM's data and AI ecosystem

- DataStage and Watson Pipelines share a common canvas built on open-source (Elyra) which is also used by services such as SPSS.
- Data integration, machine learning, data science use common platform components – allows for synergies such as shared connections.
- Extensive APIs

You can make **this**

Look like **this**



We're a leader, 17 years running

IBM named a leader in the Gartner Magic Quadrant for Data Integration Tools

Figure 1: Magic Quadrant for Data Integration Tools

Source: Gartner (August 2022)

Strengths

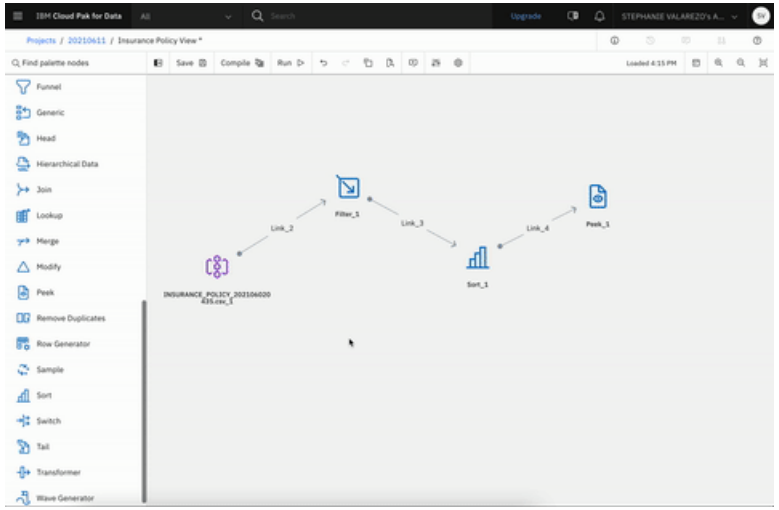
- **Support for the data fabric design:** IBM software has collaborated with IBM research to embed capabilities for augmented data integration into its Cloud Pak for Data (CPD) platform and services. The incorporation of capabilities to capture and activate metadata in Watson Knowledge Catalog, the ability to support DataOps patterns for improved orchestration and agility, and the utilization of knowledge graphs to support semantic modeling and taxonomy to ontology mapping for unstructured content have further improved its support for data fabric use cases.

- **Comprehensive portfolio for operational and analytical use-case support:** IBM has a comprehensive tools portfolio within CPD that includes DataStage (for bulk/batch integration), IBM Cloud Pak for Integration (for application integration and API management), Watson Query (for data virtualization), IBM Data Replication (for data replication and synchronization) and IBM Streams (for stream data integration scenarios). Along with these capabilities, IBM CPD is well-integrated with other data management technologies including data quality, MDM and data governance.

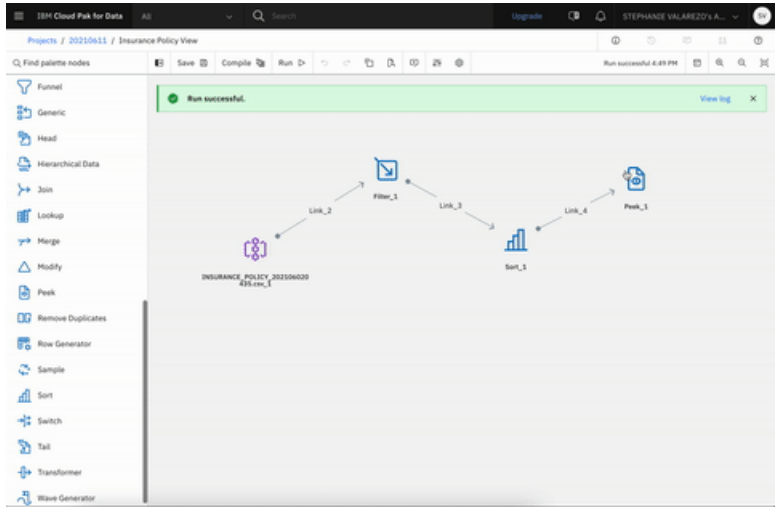
- **Modular architecture and DataOps enablement:** IBM's data integration tools are delivered as tightly integrated and yet loosely coupled services on Red Hat OpenShift (a Kubernetes-based platform). Clients praise IBM's remote runtime capabilities, which reduce egress costs by allowing developers to build pipelines once and push down workloads to the execution environments of their choice. IBM's support for CI/CD and integration with Git (for versioning), Jenkins (for task scheduling) and other third-party task and workflow managers is highly rated.

Source: Gartner, Magic Quadrant for Data Integration Tools, August 2022

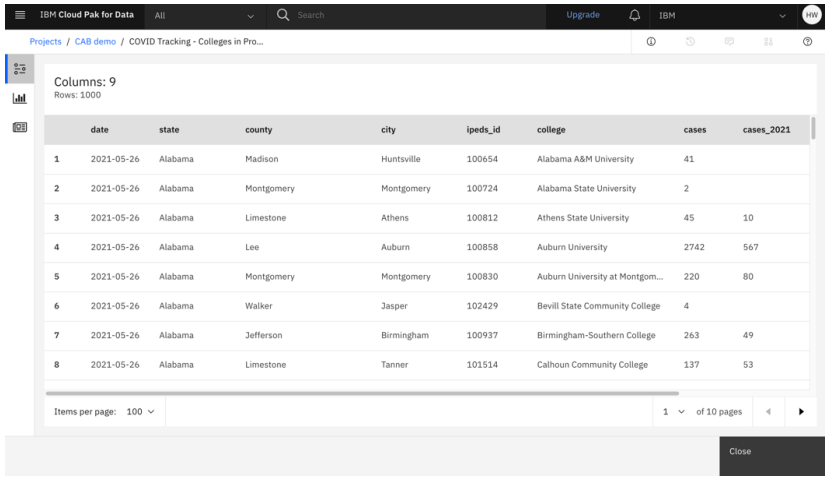
Low-code/no code visual design of data flows with hundreds of built-in transformation functions

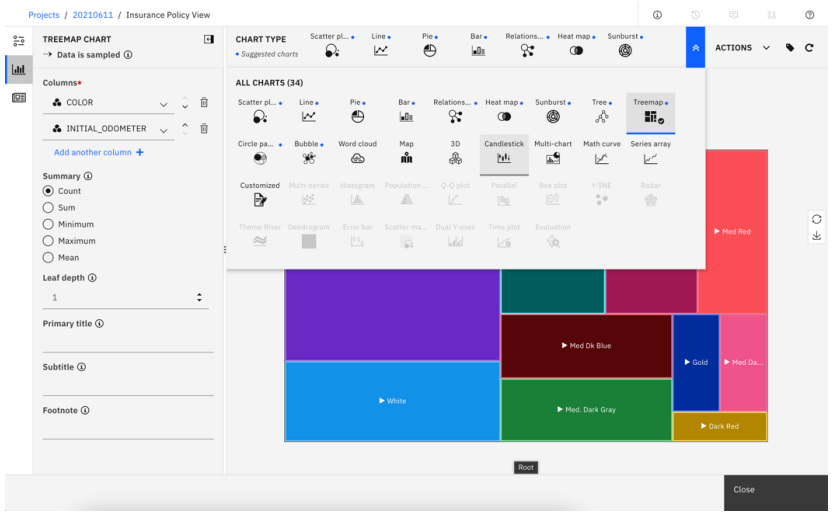


Design a job from scratch, and you will notice you do not spend a lot of time mapping columns due to **Auto-Column Propagation.** The image above demonstrates how DataStage automatically propagates a new column to downstream stages. Also, check out how we drop a stage in a link. Stages include properties developers are used to.



When you execute a flow, the job **log panel** allows developers to filter, search, and save logs directly from the designer. The links are interactive, so if you click on a stage, the canvas is re-centered on that stage.

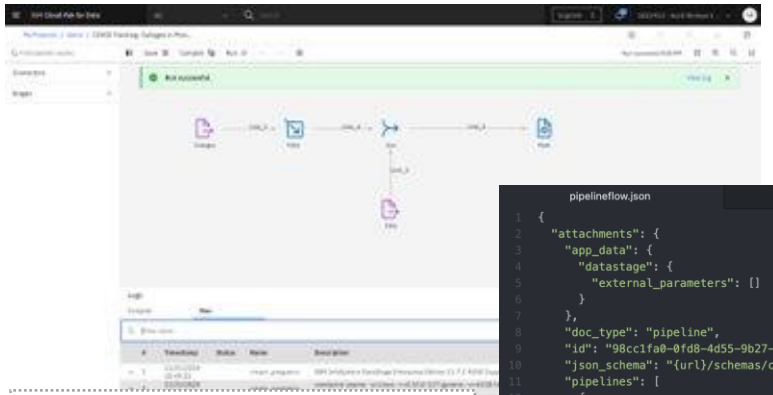




Preview data in tabular format or use the **built-in visualization** capability to quickly plot data.

Developer-centered interface with an open-source JSON format pipeline definition

- DataStage flows can be constructed in the visual interface or defined by an **open-source JSON** definition.
- API/CLI allow developers to make programmatic updates to their DataStage flows.
- https://dataplatform.cloud.ibm.com/data_intg/v3/ds_apidoc/api/explorer/#/
- https://api.dataplatform.cloud.ibm.com/v2/jobs/docs/swagger/#/



Productivity enhancements that delight developers.

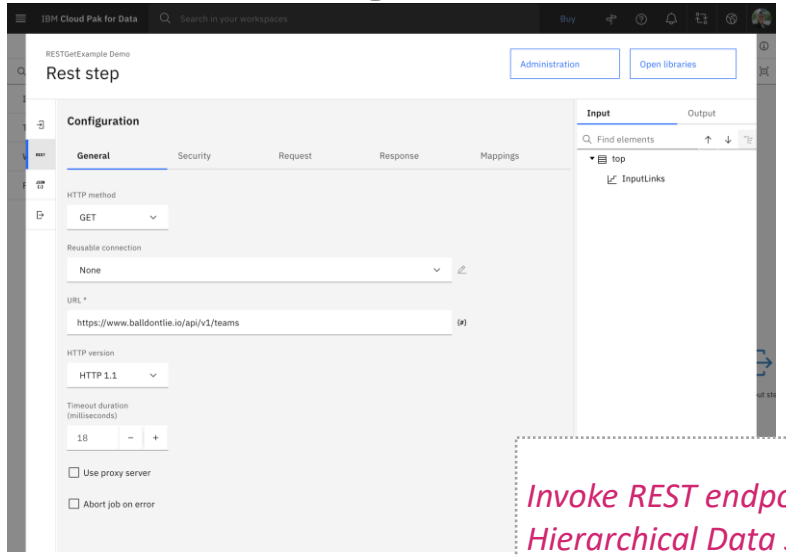
pipelineflow.json


```
{
  "attachments": {
    "app_data": {
      "datastage": {
        "external_parameters": []
      }
    },
    "doc_type": "pipeline",
    "id": "98cc1fa0-0fd8-4d55-9b27-d477096b4b37",
    "json_schema": "{url}/schemas/common-pipeline/pipeline-flow/pipeline-flow-v3-schema.json",
    "pipelines": [
      {
        "app_data": {
          "datastage": {
            "runtime_column_propagation": "false"
          },
          "ui_data": {
            "comments": []
          }
        },
        "id": "287b2b30-95ff-4cc8-b18f-92e23c464134",
        "nodes": [
          {
            "app_data": {
              "datastage": {
                "outputs_order": "46e18367-1820-4fe8-8c7c-d8badbc76aa3"
              },
              "ui_data": {
                "image": "../graphics/palette/PxRowGenerator.svg",
                "label": "RowGen_1",
                "x_pos": 239,
                "y_pos": 236
              }
            },
            "id": "77e6d535-8312-4692-8850-c129dcf921ed",
            "op": "PxRowGenerator",
            "outputs": [
              {
                "app_data": {
                  "datastage": {
                    "is_source_of_link": "55b884a7-9cfb-4e02-802b-82444ee95bb5"
                  },
                  "ui_data": {
                    "label": "outPort"
                  }
                },
                "id": "46e18367-1820-4fe8-8c7c-d8badbc76aa3",
                "parameters": {
                  "buf_free_run": 50,
                  "disk_write_inc": 1048576,
                  "max_mem_buf_size": 3145728,
                  "queue_upper_size": 0,
                  "records": 10
```

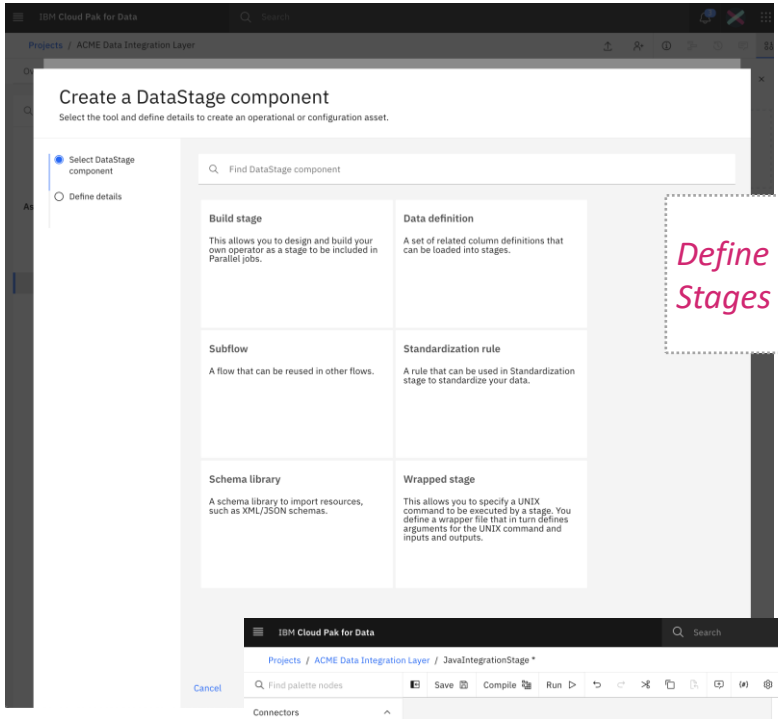

Pipelines can be constructed in the visual interface or defined with a open-source (Elyra) JSON definition.

Extensibility

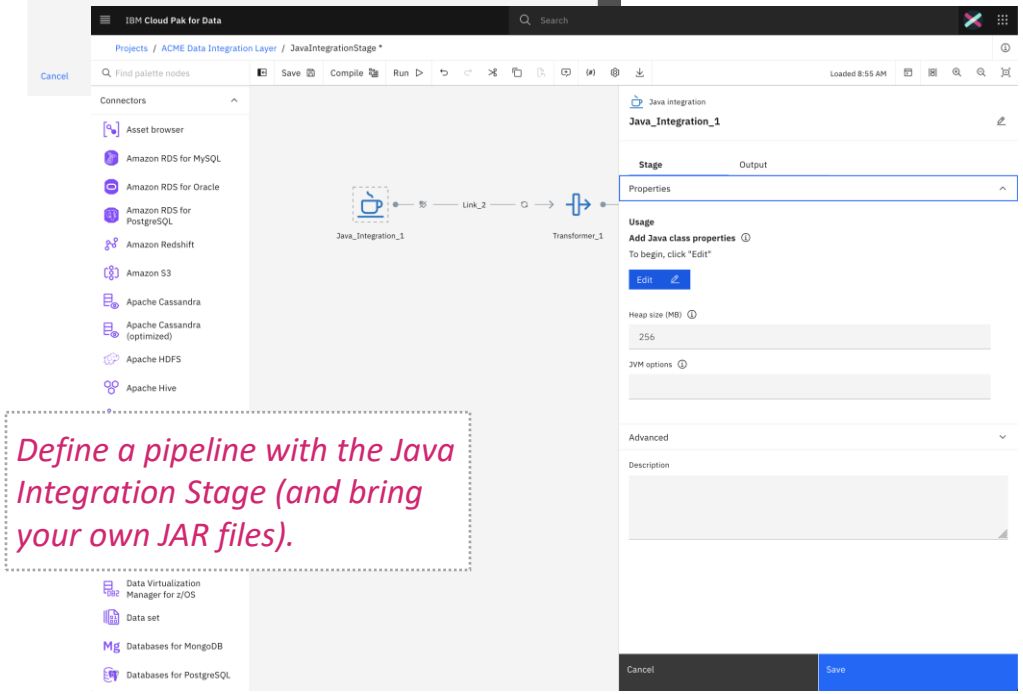
- Build your own operators based on C++; UNIX commands; bring your own Java code via the Java Integration Stage; invoke REST endpoints via the Hierarchical Data Stage.



Invoke REST endpoints via the Hierarchical Data stage.



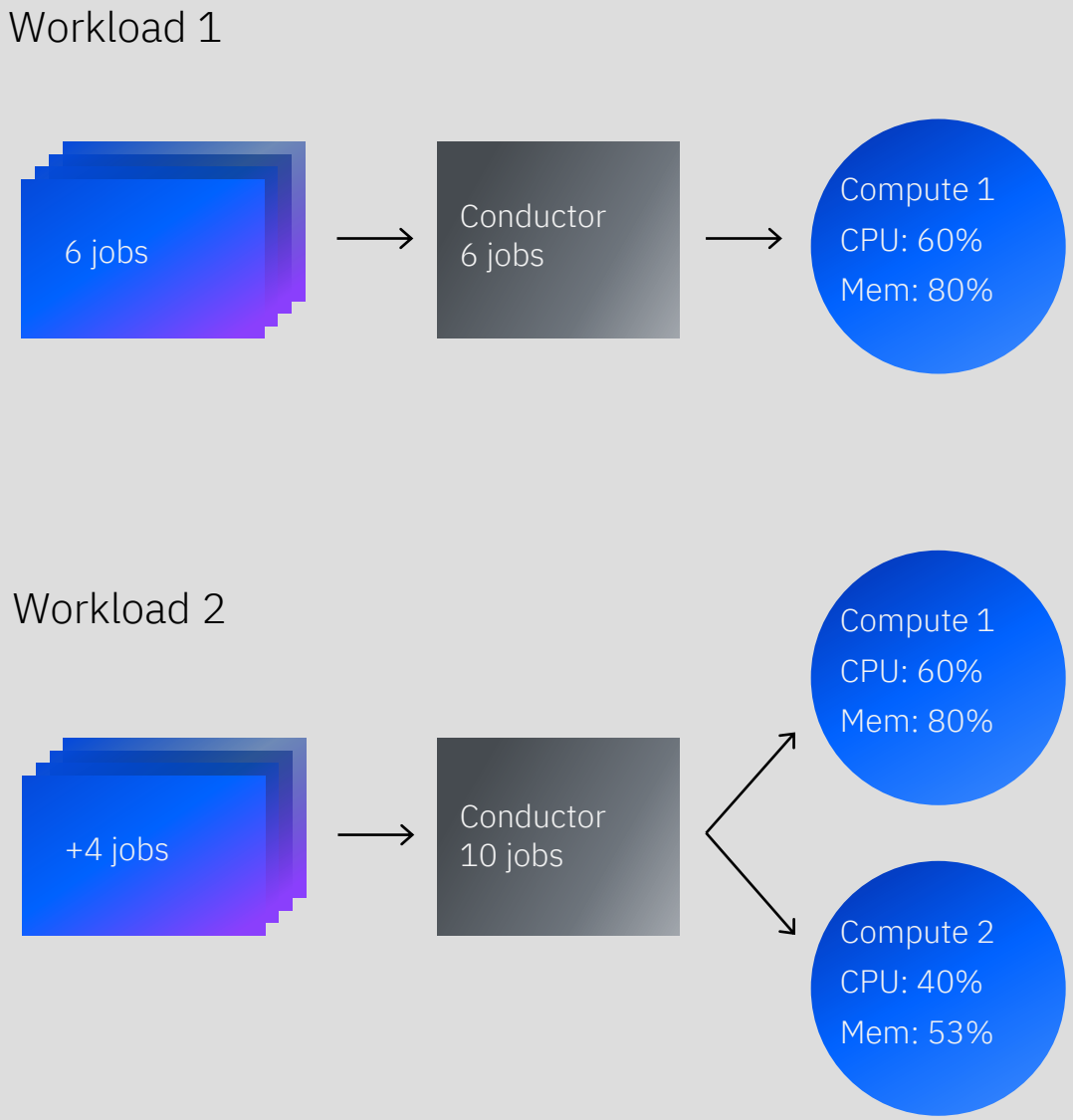
Define Build and Wrapped Stages via the user interface.



Define a pipeline with the Java Integration Stage (and bring your own JAR files).

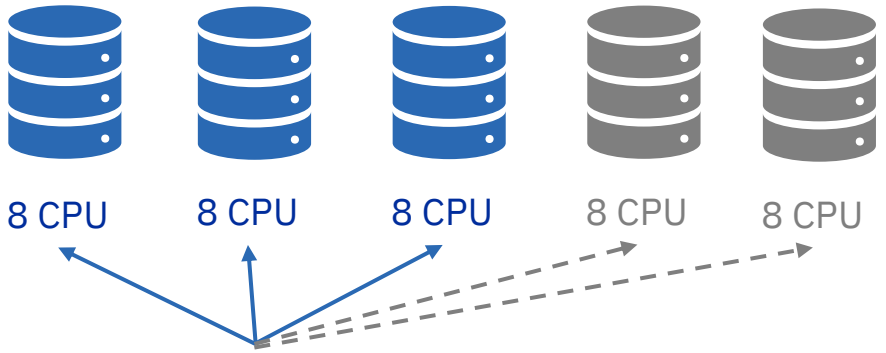
Automatic workload balancing and best-of-breed parallel engine

- Virtually unlimited scaling (horizontal, vertical) with the best-of-breed parallel engine
- Automatic load balancing maximizing throughput and minimizing resource congestion
- Automatic parallel pipelining to achieve maximum scalability and throughput
- High resiliency through automatic restart at point of failure capability
- Built on container-based architecture to allow for handling of any data volume and execution on any environment



Auto-scaling and dynamic workload management

- Dynamic workload management with Next-Gen DataStage provides **automatic workload balancing** to maximize throughput and minimize resource congestion. Jobs are managed across compute pods and queued in FIFO. If auto-scaling is enabled, then additional compute pods are deployed as needed if there is a burst in workload. And, when the workload decreases, those auto-scaled pods are scaled back down until they are needed again.
- Dynamic configuration file automatically chooses which compute pods to run jobs on at run time – no need to configure the parallel configuration file manually.



Dynamic workload management creates a parallel configuration file at runtime and automatically chooses the compute pods on which a job should run.

When auto-scaling is enabled, additional ds-compute pods are spun up based on the custom resource definition.

POLLING QUESTION 2

Current and Planned DataStage Workloads (choose as many as needed)

- On premise Data Integration
- On Premise Data Warehouse / BI
- Cloud Data Integration
- Cloud Data Warehouse / BI
- Hybrid integration / Data warehouse between on premise and cloud

DEMO of the software

PETER KEITH – IBM Brand Technical Specialist Data & AI

DataStage as-a-Service

Quick deployment

Easily provision and instantly deploy from IBM Cloud / Cloud Pak for Data catalog and use sample projects

Fully managed service

Eliminates cost and complexity of managing IT infrastructure and apps

Always up to date

Latest features and updates deployed on cloud ensuring access to the latest and greatest capabilities across the platform.

Serverless and elastic

Scale up only when needed and change configurations for runtime easily by job

Hybrid and multi-cloud

IBM managed DataStage runtime on AWS – easily run jobs on Satellite locations

Consumption-based pricing

Based on use of DataStage runtime with flexibility to scale-up when needed

POLLING QUESTION 3

Supporting Applications – Data Catalog

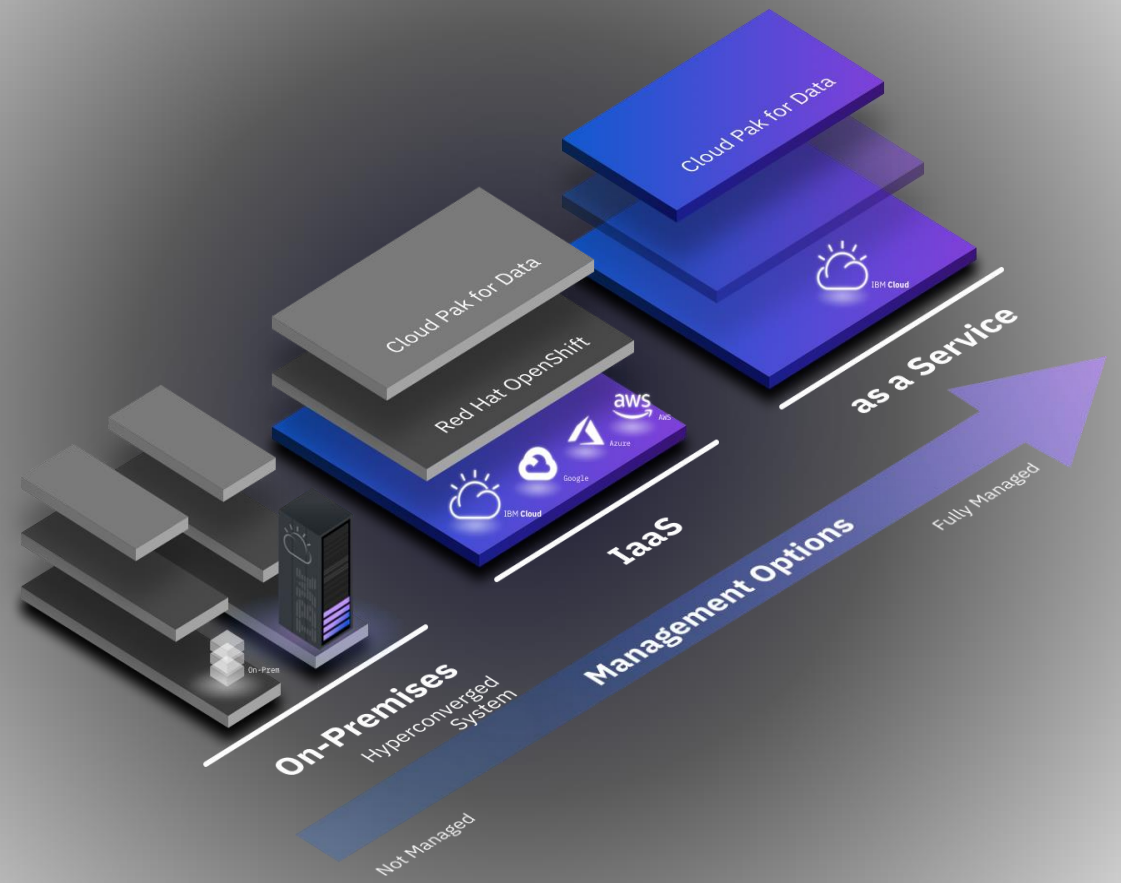
- Currently use a Data Catalog
- Actively looking for a Data Catalog now
- Planning for Catalog in the future

Flexible deployment and licensing options – from fully managed by IBM to traditional deployment options with hybrid and dual entitlements

GEMMA WOODCOCK – IBM Senior Data & AI Technical sales

DataStage – Flexible deployments

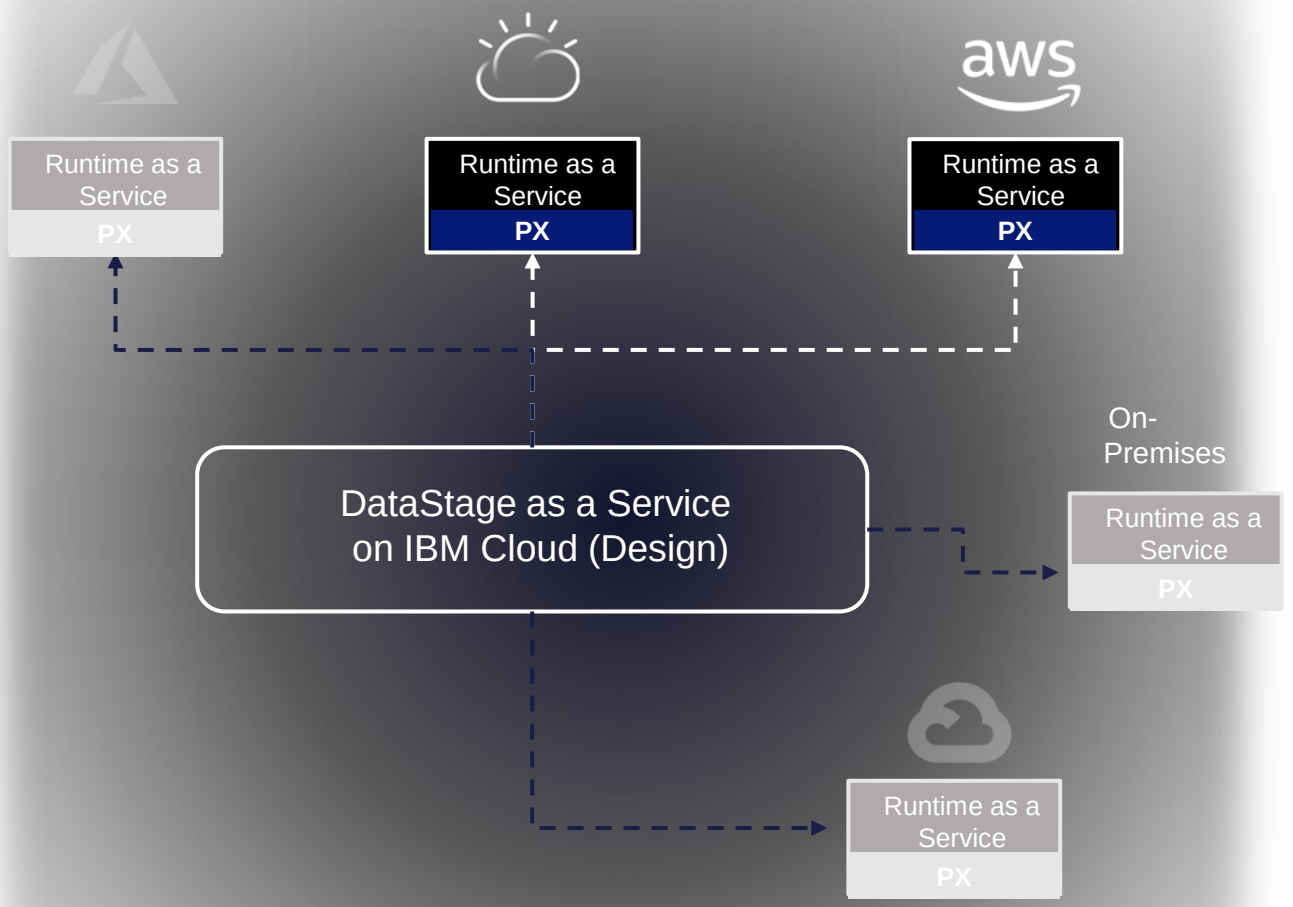
- DataStage / Information Server (stand-alone)
  - Traditional deployment on bare metal or virtual environments
  - Deploy on-premises, private cloud, or any public cloud (BYOL)
- DataStage / Information Server on Cloud Pak for Data
  - Fully containerized on a true multi cloud platform
  - Run on any cloud including on managed container service
- DataStage SaaS
  - Fully managed SaaS service
  - SaaS Runtime on other Clouds
  - Consumption based pricing



DataStage as a Service

Designed and built on cloud-native principles

- Provision the SaaS service to get started – no install or configuration needed
- No upgrades needed – fully managed
- DataStage data pipelines can be built and deployed in minutes
- New user experience with productivity enhancements for developers
- Connect to data stores you need to build DataStage flows
- Elastic scaling so you scale quickly and meet your workload requirements
- Consumption based pricing



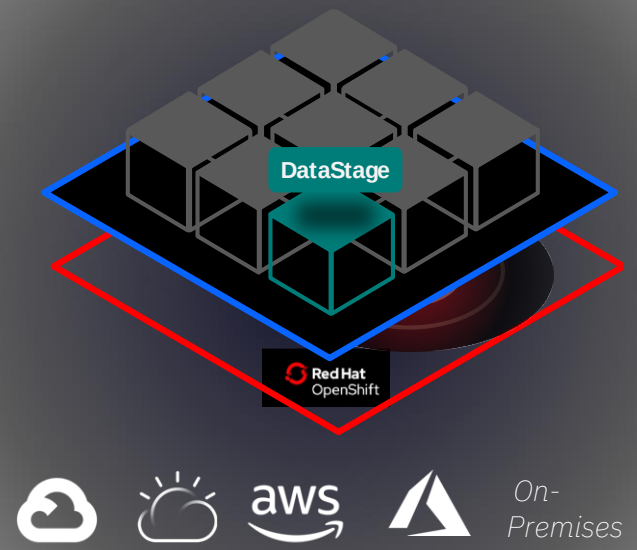
*Currently available on IBM Cloud and AWS. Other remote locations coming 2023

Accelerate time to value and reduce TCO

Flexible, elastic scaling based on workload requirements

DataStage on Cloud Pak for Data

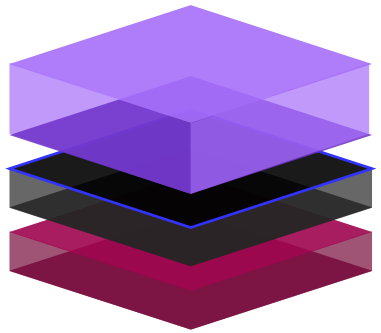
- DataStage Enterprise / Enterprise Plus
  - Fully containerized – same as SaaS
  - Tightly integrated with CP4D and other modules (esp. Watson Knowledge Catalog, Watson Studio)
  - Exploits other services (access control, projects)
- Cloud Pak for Data Control Plane
  - Common micro-services for data services
  - Authentication and roles-based authorization
  - Workload Management, Backup & Restore
  - Common UI design throughout aids collaboration, reduces product skill barriers
- RedHat OpenShift
  - Enterprise grade container orchestration
  - 100% Open Source
  - Runs anywhere



DataStage additional licensing options

New customers, or expansions (net new) for current customers

Cartridge



Containerized Offering

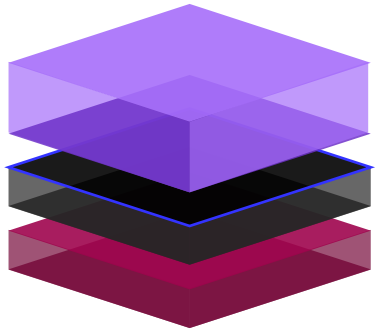
Platform Permitted Components

Red Hat OpenShift

- *Charge metric:* **Virtual Processor Core (VPC)**
- *Part types*: **Perpetual, Subscription**
- **Use entitlement for container or stand-alone**
- Control plane and other Cloud Pak for Data Enterprise Edition – permitted components only

For existing clients

Modernization



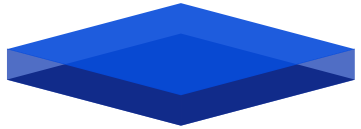
Containerized Offering

Platform permitted components

Red Hat OpenShift

Cartridge

AND



Cloud Pak for Data Enterprise Edition

- *Charge metric:* **Virtual Processor Core (VPC)**
- *Part types*: **Subscription (Upgrade), Trade-up (Perpetual)**
- **Use entitlement for container or stand-alone**
- Additional Cloud Pak for Data Enterprise Edition VPCs to run other services (e.g. Data Virtualization)

QUESTIONS & NEXT STEPS?

Some frequently asked questions

- How to migrate to DataStage as a Service
- Can you run SaaS and On Premise together
- What ‘Use Cases’ are clients using the SaaS version

For any other questions or queries please contact;

IBM - Barry O Mahoney at barry.omahoney@ie.ibm.com

OR

SmallNet – Paul Ranson at paul.ranson@smallnetconsulting.co.uk

Next Steps

- ✓Try a LITE Version of DataStage as a Service
- ✓Migrate a sample Data flow/project
- ✓Contact SmallNet
- ✓Or your IBM Account Manager

For any other help or advise

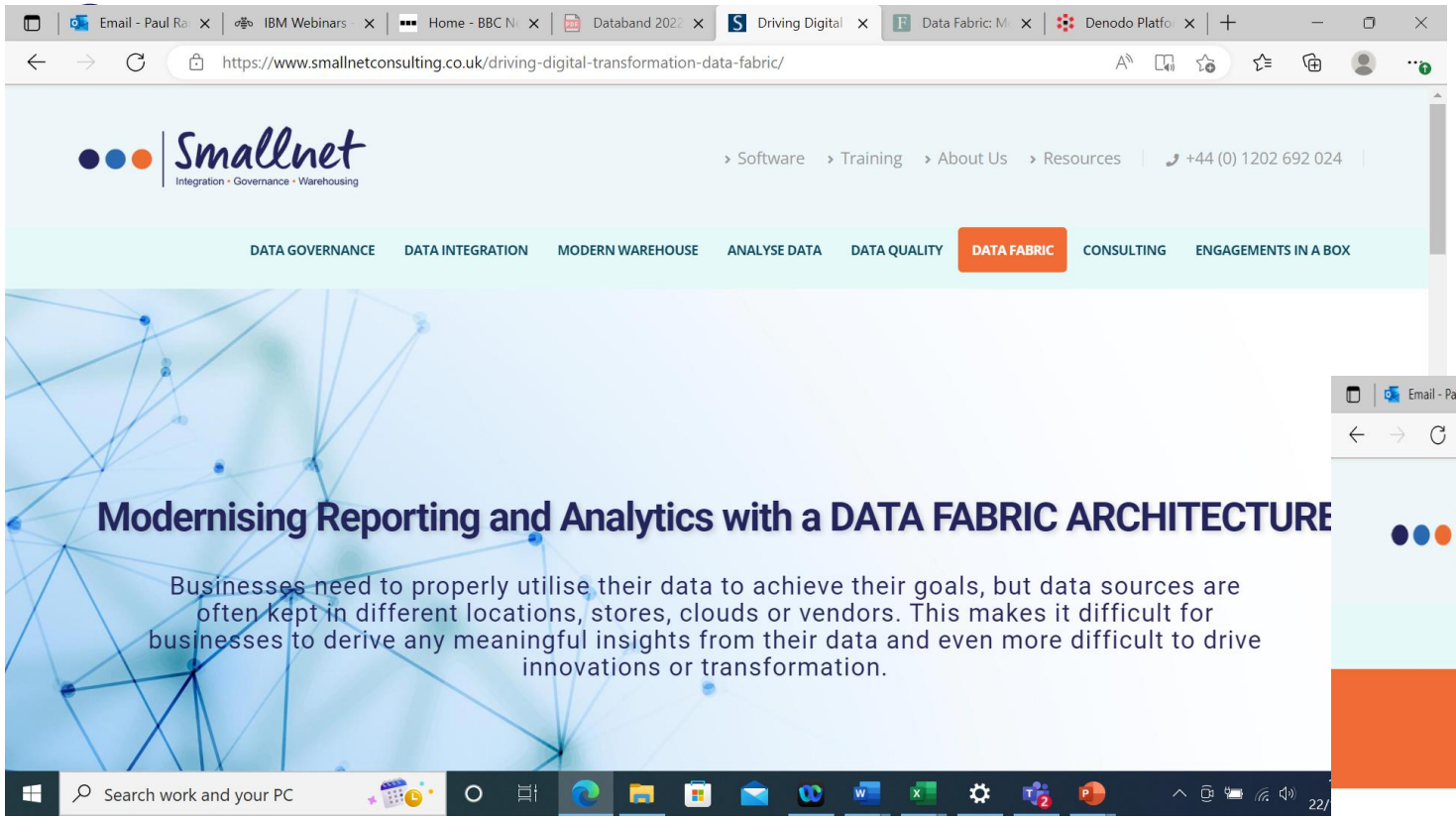
please contact;

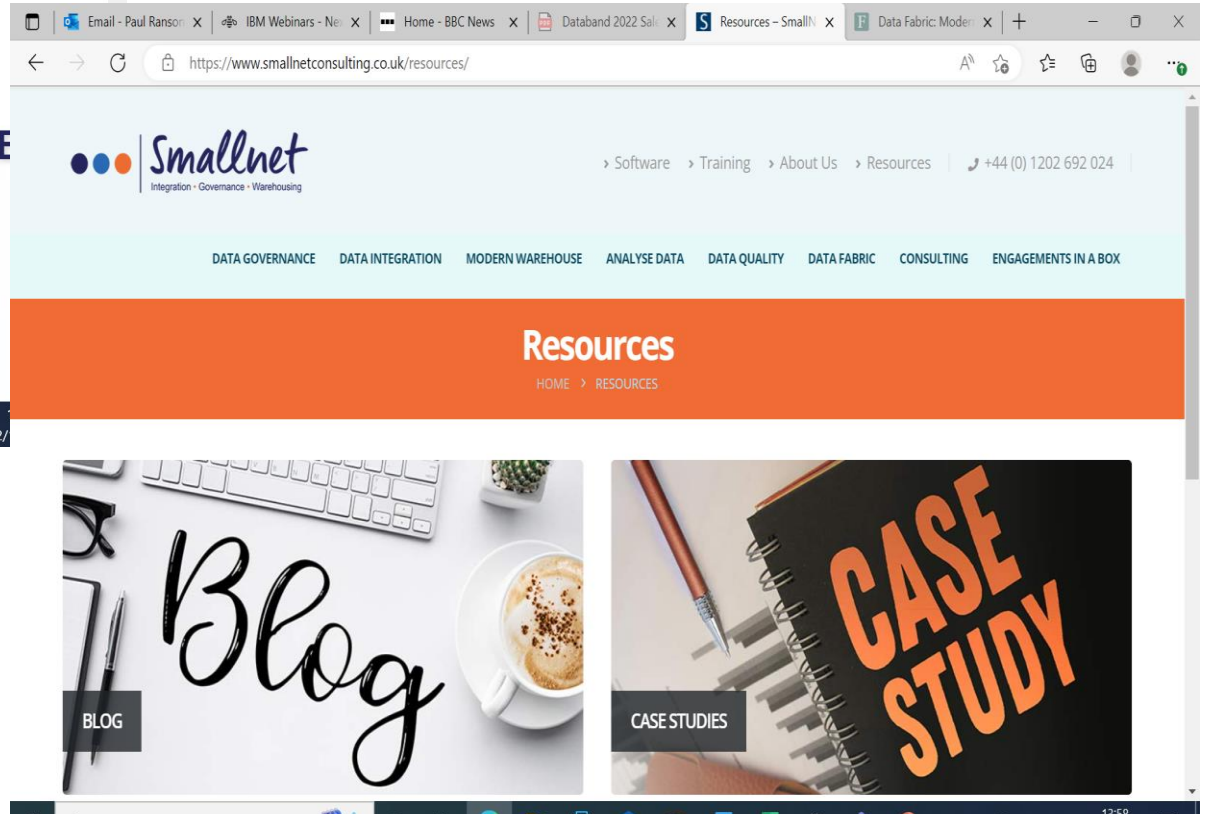
IBM - Barry O Mahoney at barry.omahoney@ie.ibm.com

OR

SmallNet – Paul Ranson at paul.ranson@smallnetconsulting.co.uk

Smallnet website contains various assets inc blogs and





Smallnet 'ENGAGEMENTS in a BOX' Accelerators and Assets

GENERAL Offerings across a wide portfolio;

- IBM Software installation and Configuration
- IBM Software Upgrade to version X to Y
- IBM Software and or DATA Health Check
- Migration from version/version or platform/platform
- Analysis and Design – Data Governance
- Analysis and Design – Data Quality
- Analysis and Design – Data Integration

IGC and WKC Modernisation Offerings;

- Planning and Implementing a Glossary on WKC / IGC
- Planning and Implementing Data Quality on WKC / IGC
- Design and Architect Data Governance on WKC / IGC
- Design and Architect Data Quality on WKC / IGC
- Migrating from IGC to WKC
- Migrating from IA to WKC

Cloud PAK for Data Solutions/Offerings;

- ✓ Getting started with DataStage on CP4D
- ✓ Getting started with Data Virtualisation on CP4D
- ✓ Getting started with Data Governance on CP4D
- ✓ Getting started with Data Quality on CP4D
- ✓ CP4D Setup and Configuration
- ✓ Migrating from DataStage to CP4D DataStage Cartridge
- ✓ CP4D Data Model Accelerators
- ✓ CP4D Data Migrations and POC's/Trials

[Check us out; https://www.smallnetconsulting.co.uk/engagements-in-a-box/](https://www.smallnetconsulting.co.uk/engagements-in-a-box/)

Useful links - materials

DataStage;

https://www.ibm.com/products/datastage

DataStage on the CLOUD, plus Trial Lite version:

https://cloud.ibm.com/catalog/services/datastage

TRY CLOUD PAK for Data Lite version;

https://eu-gb.dataplatform.cloud.ibm.com/registration/stepone?context=cpdaas&apps=all&preselect_region=true

DataStage and other Cartridges link;

https://www.ibm.com/products/cloud-pak-for-data/cartridges

Data Integration Gartner Magic Quadrant;

https://www.gartner.com/doc/reprints?id=1-2AV770E4&ct=220818&st=sb&linkId=178492211

Data Quality Gartner Magic Quadrant;

https://www.gartner.com/doc/reprints?id=1-2BKGLAS4&ct=221101&st=sb

NEXT GENERATION IBM DATASTAGE - ON CLOUD, AGILE AND AFFORDABLE

THANK YOU FOR ATTENDING

Presented by

IBM

Smallnet
Integration • Governance • Warehousing