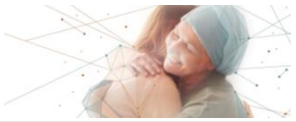


HDR UK: Data Quality Tool Assessment Request

Overview

As part of the work in HDR-UK Data Utility Workstream, we are testing several open-source/free tools to profile data quality. This will allow us to determine the impact of activities to improve the data quality.

We would sincerely like to thank you for your willingness to evaluate/run specified tools in order to provide the outputs and any feedback to HDR-UK and Inspirata.

This document provides a high-level overview and product information for the two data profiling tools you are being requested to evaluate/run against real-world data.

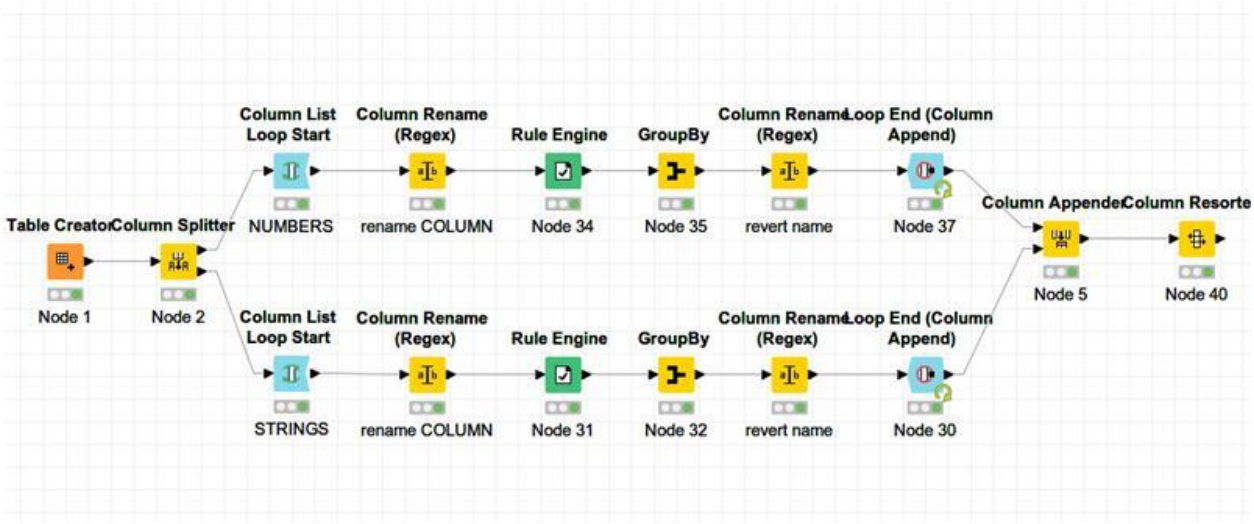
Each tool should be tested against your identified dataset(s) and evaluated in terms of ability to achieve the below mentioned criteria 1-3.

In addition, if possible, we also request that you evaluate and provide feedback on a Data Quality Monitoring/Testing tool called MobyDQ – Criteria 1 **only**. MobyDQ is a tool for data engineering teams to automate data quality checks on their data pipeline, capture data quality issues and trigger alerts in case of anomaly.

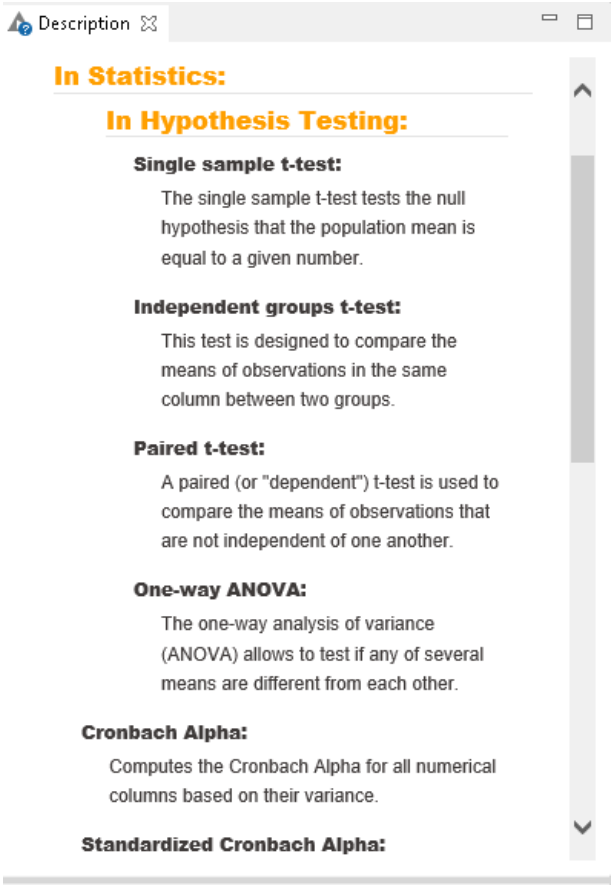
Criteria 1: Please provide any feedback on your experience of the tool

Please provide any feedback, screenshots and comments that you believe may be pertinent.

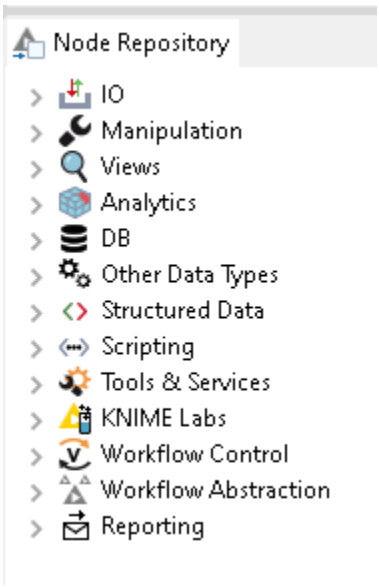
- KNIME is quiet resource intensive for system resources, it took a while to interrogate data with i5 Coffeelake processor overclocked at 4.2ghz and 16gb of ram.
- We store our verified data sets in .dta files for Stata and therefore had to convert to .csv for use in KNIME.



- Complex workflows could be created and run, however it took substantial time to get to grips with the tool.
- The node repository seemed vast with detailed explanations on how to work a node into the workflow. However as previously mentioned complex workflows were hard to get to grips with and lots of time had to be spent debugging and learning the basics.



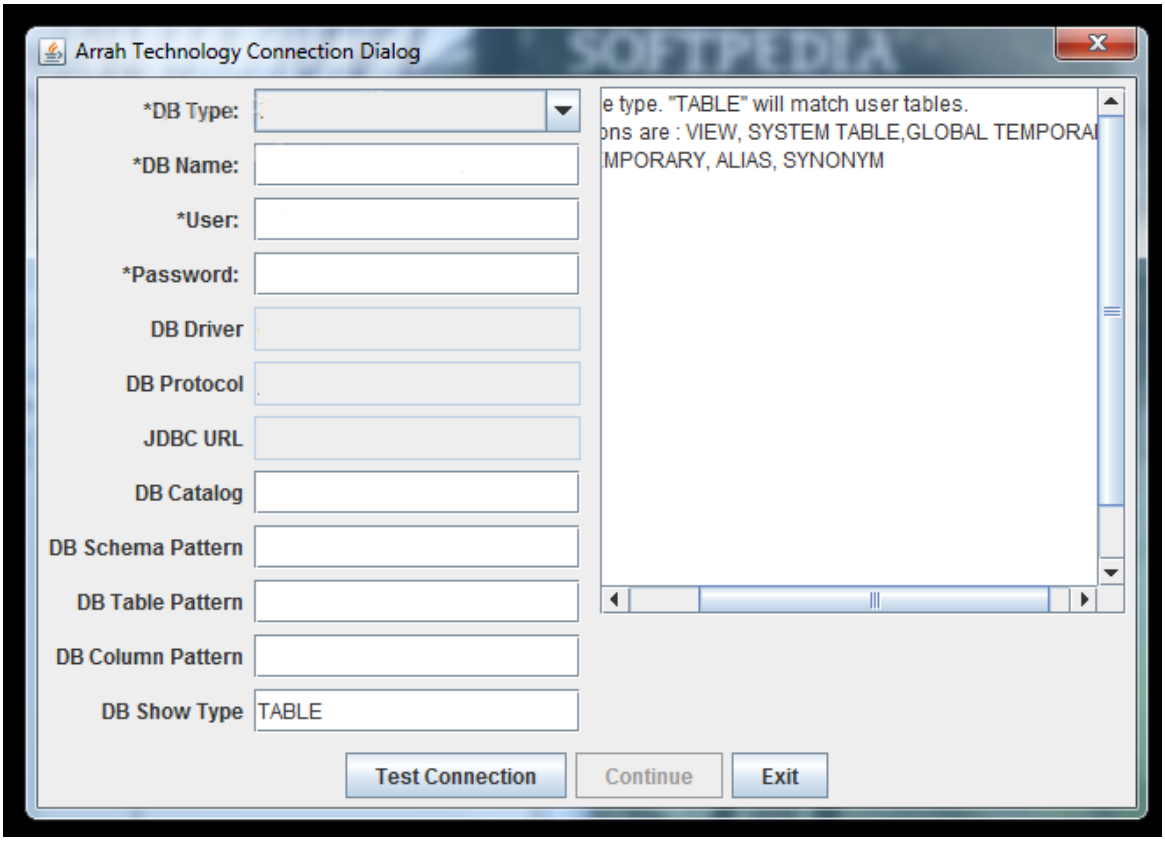
- The in window explanations (show above) were very detailed and helpful, and given the days and weeks it would take to learn the tool im sure it would effectively interrogate most data.
- The huge number of extensions and functionality meant that when searching for soloutions I was given multiple routes to achieve my goal, which can be both an advantage and a disadvantage
- KNIME had the abilty to function with R coding, our stats team primarily use STATA.



-
- The node repository was vast with 100s of tools to intergrate into workflows. With adequate training KNIME looks like it could handle huge data sets.

Aggregate profiler

- The aggregate profiler presented difficulties from the start, after installation and running of the windows bat file I was presented with a command prompt window, and the arrah connection dialog. I attempted to fill out the connection dialog and then was prompted to update my Java, upon updating my JAVA the system crashed and would not reboot.

- 

- Additionally the Aggregate profiler required specific knowhow/qualifications that I did not meet – "Window User should know how to create DSN or Use existing DSN for MS Access, SQL Server, MySQL, and Oracle on Windows. "

Criteria 2: Can the Tool profile/calculate the following?

Assessment values:

- Yes – Tool can do this
- No – Tool cannot do this
- Partial – Tool can partially do this

Category	#	Feature	Knime	Aggregate Profiler
Single Column – Cardinalities	1	Number of rows	YES	
	2	Number of nulls	YES	
	3	Percentage of nulls	YES	
	4	Number of distinct values (cardinality)	YES	
	5	Percentage of distinct values (Number of distinct values divided by the number of rows)	YES	
	6	Frequency histograms (equi-width, equi-depth, etc.)		

Single Column - Value distributions	7	Minimum and maximum values in a numeric column		
	8	Constancy (Frequency of most frequent value divided by number of rows)		
	9	Quartiles (3 points that divide the **numeric** values into 4 equal groups)		
	10	Distribution of first digit in numeric values (to check Benford's law)		
Single Column - Patterns, datatypes, and domains	11	Basic types (e.g. numeric, alphanumeric, date, time)		
	12	DBMS-specific data type (e.g. varchar, timestamp)		
	13	Measurement of Value length (minimum, maximum, average, median)		
	14	Maximum number of digits in numeric values		
	15	Maximum number of decimals in numeric values		
	16	Histogram of value patterns (Aa9...)		
	17	Generic semantic data type (e.g. code, date/time, quantity, identifier)		
	18	Semantic domain (e.g. credit card, first name, city)		
Dependencies	19	Unique column combinations (UCCs) (key discovery)		
	20	Relaxed unique column combinations		
	21	Inclusion dependencies (INDs) (foreign key discovery)		
	22	Relaxed inclusion dependencies		
	23	Functional dependencies		
	24	Conditional functional dependencies		
Advanced Multi Column profiling	25	Correlation analysis		
	26	Association rule mining		
	27	Cluster analysis		
	28	Outlier detection		
	29	Exact duplicate tuple detection		
	30	Relaxed duplicate tuple detection		

Criteria 3: How easily can the tool profile the data quality dimension measure criteria?

Ratings range from 0.0 to 5.0:

0.0 = Tool cannot calculate this
1.0 = Poor: most or all defined requirements not achieved
2.0 = Fair: some requirements not achieved
3.0 = Good: meets requirements
4.0 = Excellent: meets or exceeds some requirements
5.0 = Outstanding: significantly exceeds requirements

DAMA Dimension	Measure (key elements)	Knime	Aggregate Profiler
Completeness	Percentage of requisite information available	3	
	Percentage of missing data values (null / empty string)	3	
	Row counts	3	
	Highest and lowest value of key elements	2.8	
	Number of data values in an unusable state	3	
Uniqueness	(Number of things in the real world) Number of incorrect spellings etc. of same data in an element e.g. address (duplicate values)		
	(Number of recodes describing different things) Number of data items in adherence to expected/described data element value (distinct values at ID level)		
	(Number of things in real world i.e. duplicates)/ (Number of records describing different things i.e. distinct records)		
Timeliness	Difference between Lowest date value and Highest Date Value		
	Number of records per month		
Validity	Percentage of data values that comply with the specified formats (data types, ranges etc.)		
	Percentage of data values that don't comply to specified formats		
	Number of Missing values indicated e.g. with fill values		
	Number of Values in Specified Range		
	Number of values not in Specified Range		
Accuracy	Number of accurate data values		
	Number of inaccurate data values		
	Actual data value count versus predicted data value count		
	Number of rows and columns against expectations		
	Number of duplicates at ID level		
	Number of blank columns, large % of blank data, high % of same data		
	Distribution across various segments		
	Outliers on key variables		
	((Count of accurate objects)/ (Count of accurate objects + Counts of inaccurate objects)		
Consistency	Analysis of pattern and/or value frequency		

Tools to Evaluate

1. Knime

Link(s):	• https://www.knime.com/
License:	GNU General Public License version 3.0 (GPLv3)
Version:	4.1.2
OS:	Windows, macOS, Linux
Description:	

KNIME Analytics Platform is an open solution for data-driven innovation, designed for discovering the potential hidden in data, mining for fresh insights, or predicting new futures. Organizations can take their collaboration, productivity and performance to the next level with a robust range of commercial extensions to our open source platform.

KNIME Analytics Platform provides the tools to connect to a host of databases and data warehouses, access a variety of file formats, retrieve data from cloud resources or external services, and more. The broad set of out-of-the-box functionality, allows you to seamlessly integrate and transform the data in one uniform, visual environment on your own - no dependencies on central IT. If there's a functionality you're missing, simply integrate the tools you like or take advantage of the many integrations we have with other open source projects. Workflows created with KNIME Analytics Platform automatically document each step of your data wrangling process. Meaning, if you share workflows or results with your colleagues, they can easily understand the individual steps of your workflow and provide feedback.

KNIME Software covers all kinds of data analytics functionality - for example classification, regression, dimension reduction, or clustering, using advanced algorithms including deep learning, tree-based methods, and logistic regression. Among these, are integrations with other large, open source projects such as Keras or Tensorflow for deep learning, H2O for high performance machine learning, R and Python for coding, and various implementations for model interpretability and validation.

From integrations with Apache Spark for big data processing, to KNIME Server distributed executors for handling concurrent workflow execution, KNIME Software ensures data science is created and deployed quickly and efficiently.

PRODUCT FEATURES:

- Powerful Analytics
- Data and Tool Blending
- Over 1000 modules and growing
- Connectors for all major file formats and databases
- Supports multiple data types: XML, JSON, images, documents etc.
- Native and in-database data blending and transformation

- Math and statistical functions
- Advanced predictive and machine learning algorithms
- Workflow control
- Tool blending for Python, R, SQL, Java, Weka and others
- Interactive data views and reporting

2. Aggregate Profiler

Link(s):	• https://sourceforge.net/projects/dataquality/ • http://www.arrahtech.com/docs/installation_guide.html
License:	GNU General Public License version 3.0 (GPLv3)
Version:	v6.3.0
OS:	Windows, macOS, Linux

Description:

Also known as Open Source Data Quality and Profiling. Features MySQL, Oracle, Postgres, Access, Db2, SQL Server certified Big data support - HIVE Format Creation, Format Matching (Phone, Date, String and Number), Format standardization Fuzzy Logic based similarity check, Cardinality check between tables and files Export and import from XML, XLS or CSV format, PDF export File Analysis, Regex search, Standardization, DB search Complete DB Scan, SQL interface, Data Dictionary, Schema Comparison Statistical Analysis, Reporting (dimension and measure based), Ad Hoc reports and Analytics Pattern Matching , Deduplication, Case matching, Basket Analysis, Distribution Chart Data generation and Data masking features Meta Data Information, Reverse engineering of Data Model Timeliness analysis , String length analysis Address Correction, Single View of Customer, Product, Golden merge for records Record Match, Linkage and Merge added based on fuzzy logic

Features:

- -Data profiling, filtering, and governance
- -Similarity checks
- -Data enrichment
- -Real time alerting for data issues or changes
- -Basket analysis with bubble chart validation
- -Single customer view
- -Dummy data creation
- -Metadata discovery
- -Anomaly discovery and data cleansing tool
- -Hadoop integration

3. MobyDQ

Link(s):	• https://github.com/ubisoftinc/mobydq • https://ubisoftinc.github.io/mobydq/ • https://ubisoftinc.github.io/mobydq/pages/productiondeployment/
License:	Apache License 2.0
Version:	1.0
OS:	Windows, Linux
Description:	

MobyDQ is a tool for data engineering teams to automate data quality checks on their data pipeline, capture data quality issues and trigger alerts in case of anomaly, regardless of the data sources they use.

Free and open-source Data Quality solution that aims to automate Data Quality checks during data processing, storing Data Quality measurements and metric results, and triggering alerts in case of anomaly. The framework can be used to access different data sources. It installed quickly and straightforward, based on the detailed documentation provided on GitHub. MobyDQ does not provide any data profiling functionality, because its focus is on the creation, application and automation of Data Quality checks.

MobyDQ provides a toolbox for data engineering teams to design data quality indicators with the objective to answer the following questions:

- Is all the necessary data present in the system?
- Is the data available at the time needed for its usage?
- Is the data compliant with validation or business rules?
- Does the data reflect real world objects?

These questions can be answered using the following types of indicators:

- Anomaly detection: Machine learning algorithms to detect outlying values. **Work in progress**.
- Completeness: Difference in percentage between a measure computed in the source system and the same measure computed in the target system.
- Freshness: Difference in minutes between the current timestamp and the last updated timestamp in the target system.
- Latency: Difference in minutes between the last updated timestamp in the source system and the last updated timestamp in the target system.
- Validity: Any measure computed in target system which does not comply with a validation or business rules.

Inspirata Overview

Inspirata is a revolutionary healthcare IT company, using cutting edge technology to help organizations and patients realize more value from healthcare data. Our goal is to transform care by empowering health systems and providers with the real-time, system-wide data and intelligence needed to improve care quality and the patient experience, and ultimately, the economics of health and wellness.

Inspirata combines the power of an open technology platform, natural language processing and AI, and collaborative clinical applications to bring together disparate patient data and transform it into intelligence. We enable our customers and partners to take advantage of the platform to innovate—rapidly generating a new era of applications to improve healthcare and deliver superior outcomes to patients. Our goal is to make it easy for caregivers across the entire healthcare continuum to gain the insight they need to collaborate and provide the best patient care possible.

Inspirata's Clinical Data Platform provides healthcare institutions the ability to bring all their data together into one place. From this megastore, numerous applications can be built, such as: analytical and data mining engines, customized workflow management, and simplified sharing of critical information with physicians and other healthcare providers worldwide.

Our industry-leading oncology specific Natural Language Processing (NLP) technology provides value by automating manual data curation processes across the oncology service line including cancer case-finding, cancer reporting, clinical trial candidate identification and quality monitoring.

Inspirata's solutions enable our customers to achieve efficiencies, delivery higher quality care, reduce costs, engage in ground-breaking research programs, and participate in partnerships with pharma/biotech companies.