THE DATA LAB
value from data

HIE
Highlands and Islands Enterprise
Iomairt na Gàidhealtachd 's nan Eilean

Scottish Enterprise

Data Landscape

Factsheet

DEMYSTIFYING THE LANGUAGE OF DATA

Data of itself has little value, however when it is analysed it is transformed into information and it is only when action is taken, informed by the insight that the information delivers, that the value of data is realised.

For some, being able to do this conjures up the need for complex IT systems and expert data scientists developing complex data strategies. Indeed, for some this will be true, but there are lots of ways in which businesses of any size can utilise data to improve their productivity and competitiveness. Do not be fall into the trap of thinking that this is too difficult or too expensive and be put off considering how data can benefit you.

This factsheet explores some of the terminology you might come across associated with data and seeks to demystify the jargon that is commonly used. It also introduces a step-by-step process of how you would go about using data and identifies some of the technologies and tools involved.

AN APPROACH TO EFFECTIVELY USING DATA – STEP BY STEP

1. Questions, questions...

- The start of the process is about identifying the business objective:
- What do you want to achieve from the analysis of your data?
- Is it further insight into how your customers behave?
- What savings can be made?
- How can the business anticipate and solve a problem before it arises?

2. Data audit

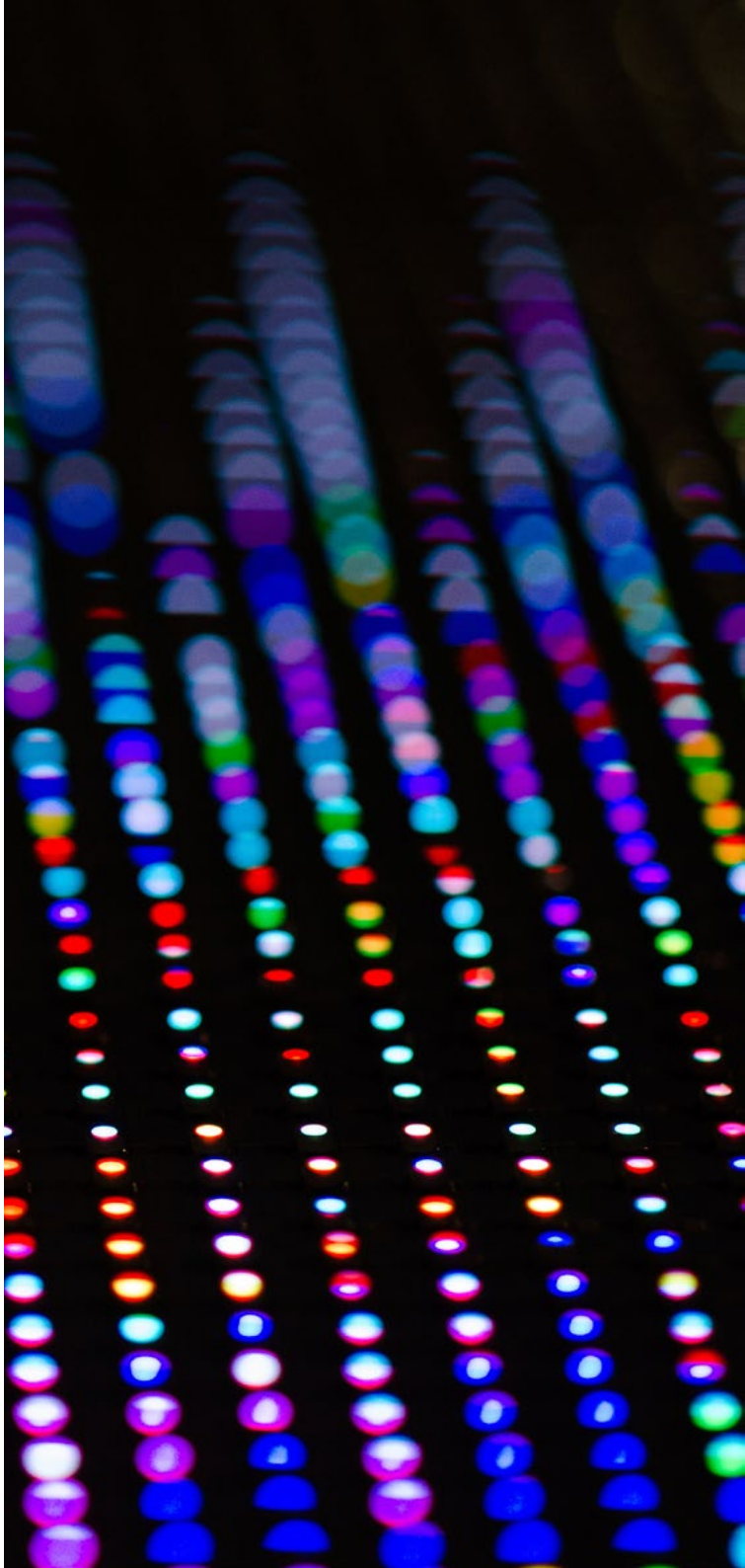
A data audit is the process of examining data - assessing its quality and potential use for a specific purpose. Auditing data involves looking at key metrics to understand the properties of a data set and understanding what data the business has, who collects that data, how it is used, who uses it, if it is accessible and who would benefit from it. It should also assess ownership, dependency, security, quality and also what data (if available) would be useful to the business. A data audit reviews the origin, creation, or format of data to assess its value and potential uses.

3. Data acquisition

This is about data collection and storage. It is about identifying and accessing the data that is relevant to your objective (whether internal or external data) and understanding its properties. As we could be talking about a huge number of digital files in various formats, where this will be stored is a key consideration. By collecting and ordering your data appropriately, it can then be retrieved and analysed effectively.

At this stage you will hear phrases such as:

- **Data sets:** a collection of information organised into data structures such as an Excel document, a table, a file, an information tree, etc.
- **APIs (application programme interfaces):** sets of routines, protocols, and tools that allow data to be exchanged between computer programmes or systems and allow you to build software applications to access that data. These can be used to access key sources of data in the internet economy and enable new services to be created.
- **Open data:** data that is free and publically available for anyone to use. It is not restricted by patents or copyrights, but must have a license to say it is open data. Though this is not your own data, it can be used to provide further insight similar to buying data.



- **Buying data/data brokerage:** other datasets generated from businesses or public bodies can, in some circumstances, be purchased to supplement your business' existing data and provide further insights.
- **Data cleansing**: an examination of the data that has been identified and removing duplicate, inaccurate, corrupt, incomplete, or irrelevant information. The data is then structured into a data set, which is consistent with other data sets in your system. In practice this could be as simple as correcting a typo in an Excel document or removing an address in a database that doesn't have a postcode.
- **Web scraping:** the process of extracting large amounts of information from a website and saving it locally in a file on a computer, database or spreadsheet. Often this data is only viewed through a web browser without the ability to download it; for example, a website directory. Therefore, it must be manually input into a spreadsheet. Web scraping is the automation of this process using tools and software to cut out the human element of this task.

4. Data analytics

This is the application of data science to extract valuable information from the raw data collected in the first stage and where data hypotheses are developed and patterns and anomalies are identified. Analysis is undertaken using some of the data technologies/ applications/tools listed below.

5. Data visualisation

This is a process where the outcomes of data analyses are presented in an easy to understand format. Data is displayed as graphs or images as a way of explaining the results rather than an Excel spreadsheet of impenetrable facts and figures.

KEY DATA SCIENCE BUZZWORDS

- **Algorithm:** a step-by-step process designed for a computer to follow in order to solve a problem.
- **Artificial intelligence (AI):** the intelligence of machines. In data science this is a machine understanding and perceiving its environment. Based on these perceptions it can then make decisions to reach an outcome.
- **Machine learning:** the construction and analysis of algorithms for a computer to learn and make predictions based on data. An analysis of this data means machines can perform basic tasks, like facial recognition in photos.
  - **Deep learning (a.k.a. deep structured learning, hierarchical learning or deep machine learning):** a branch of machine learning that seeks to model high level abstractions in data. For instance – automatic image captioning, where the subject of an image is recognised and captioned appropriately.
- **Predictive analytics:** the analysis of historical data to help determine patterns and predict future outcomes and trends.
- **Internet of Things (IoT):** the interconnectivity of objects. Interconnection of so called 'smart' objects is enabling the roll-out of developments such as smart grids and smart cities. The predicted proliferation of billions of connected devices will massively increase the data we have available with almost every aspect of our environment monitored and accessible.
- **Data architecture:** A broad term describing the application of data science in an organisation. The implementation of a set of standards and models that govern and define the type of data collected and how it is used, stored, managed and integrated within an organisation and its database systems. Major players in the data architecture field are IBM, Cloudera, MapR, and HortonWorks.

DATA TECHNOLOGIES, TOOLS AND APPLICATIONS

In creating your data architecture it may be built on a platform of several standard software applications and may use a range of tools that deliver specific tasks. Here are some of those that are commonly used and discussed in this context:

- **Data technologies/applications/tools:** These terms are interchangeable where tools and programmes are used to deliver data architecture. A data application is a computer programme whose primary purpose is entering and retrieving information from a computerised database:
- **Cloud:** when data is stored on remote computers accessed over the internet as opposed to on a local computer.
- **Data lake:** where data is stored in an unstructured format (i.e. before the data has been cleaned or organised and placed into a data warehouse).
- **Data warehouse:** where data is stored in a structured format (i.e. it has been cleaned and put into appropriate or relevant data sets).
- **RDBMS (relational database management systems):** a database management system. RDBMSs have been used since the 1980s and are commonly used for storing structured information in databases used for financial records, manufacturing and logistical information, and personnel data. In its simplest form, it consists of a collection of tables with each consisting of a set of rows and columns.
- **SQL** (pronounced "sequel"): Structured Query Language is a tool or language used for querying and managing relational databases.
- **Open source software:** software that is free and open to use by the public and may be developed in a collaborative public manner. The Apache Software Foundation provides support for the Apache Community of open source software projects such as Apache Hadoop (see below).
- **Apache Hadoop:** a big-data framework. It distributes massive data collections as well as indexes and tracks this data across a large cluster of computers, enabling big-data processing and analytics far more effectively than was possible previously.
- **Apache Spark:** another big-data, open-source framework developed specifically for handling large-scale data processing and analytics. The difference between Hadoop and Apache Spark is that the latter is generally considered faster.
- **Python:** is a general purpose data analytics programming language. Software programmes are written in code and Python enables these programmes to be interpreted.
- **MATLAB (matrix laboratory):** software for technical computing. It integrates calculation, visualisation, and programming in an easy-to-use setting where problems and solutions are expressed in familiar mathematical notations. Typical uses include: maths and computation and algorithm development.
- **Scilab:** a high-level programming software, but it is numerically oriented and open-source. Scilab is commonly used for engineering and scientific applications.
- **GNU Octave:** another high-level programming software also focused on numerical calculations. Octave helps in solving linear and nonlinear problems numerically.
- **R:** a programming software for statistical computing and graphics. R is typically used by quantitative analysts in the financial sector.
- **Google Analytics:** Google's "freemium" data analytics platform which can be used by any business to analyse and visualise web and mobile app traffic and is used by about half of all websites.