Power Query and Big Data

What is Big Data? We usually talk about Big Data when traditional data processing methods can no longer cope with the volume, velocity, variety and reliability of the data. While there is no specific amount of data that can be labelled as Big Data, it is usually a data set that exceeds the capabilities of traditional databases and data processing tools. What should you do if you need to analyse such a huge data set and don't have immediate access to a fully-fledged Big Data platform? In this article, we will look at how you can use available tools to gain insight into this Big Data when no specialised solution is yet available.

One such tool is Power Query. This is a powerful data transformation tool that integrates into Excel and Power BI and provides an efficient way to gain insights into Big Data. It allows you to perform initial data exploration and preparation before performing more complex analyses. Whilst Power Query is not a complete Big Data solution in itself, it is a valuable aid to understanding and preparing data for further processing

Power Query plays a crucial role in the early stages of your workflow before you can process your data. It acts as a bridge between different data sources. Whether your data is in large CSV files, large databases or elsewhere, Power Query can make the connection and extract the information you need. However, it is often impractical to load millions or billions of rows directly into Excel. This is where Power Query’s ability to analyse different data sets comes into play.

It is possible to obtain smaller, representative samples of Big Data for analysis using different methods. Obtaining the first N rows provides quick insights, while selecting a random sample provides a more statistically sound set. Filtering by specific criteria, such as transactions from a particular region or period, allows one to focus on relevant subsets of the data. Once a manageable sample of data has been selected, you can utilise Power Query’s data preparation functions. Common tasks include dealing with missing values, standardising data formats (dates, currencies, etc.), aggregating data to a higher level of granularity and creating calculated columns to gain new insights. For example, it is possible to confirm the proportion of sales by customer based on transaction history or categorise products based on sales volume.

However, it is also important to mention the limitations of Power Query when working with datasets that can be labelled as Big Data. Performance can degrade significantly as the volume of data increases, and if you are using Excel as the final repository for data, the limited number of rows can become a hindrance. Therefore, it is very important to optimise the data processing steps and queries that affect the processing speed. It is necessary to filter the data at the beginning of the process to reduce the amount of data loaded. It is desirable to reduce the number of transformation steps and use appropriate data type conversions to avoid unnecessary utilisation of processing power. This approach is suitable for initial research work and analyses on a smaller scale. For tasks that require the processing of large amounts of data, specialised Big Data platforms such as Hadoop, Spark or cloud data warehouses are required.

An example of when initial data research should be carried out is analysing customer transaction data from an e-commerce platform. With Power Query, you can connect to a database that contains transaction data. From this, you can select a sample of 10,000 transactions and then use Power Query to calculate the average transaction value per customer. Identify the top-selling product types and segment customers based on transaction history. This initial analysis can provide valuable insights and indicate which data sets are worth analysing in more detail using specialised tools.

To summarise the above, Power Query is a practical and accessible tool for gaining small insights into Big

Data. It provides data selection and processing that allows users to explore huge data sets and gain initial insights without the need for a full Big Data infrastructure. Even if no Big Data processing platform is available, you can use Power Query to perform an initial small data exploration. To expand your knowledge of Big Data, we recommend that you explore resources on data warehousing, cloud computing and distributed processing techniques.