LEARNING MADE EASY

Zetaris Special Edition

The Networked Data Platform

for dummies®
A Wiley Brand

Combine your data sources instantly

Analyse large, complex datasets in real time

Test-drive the platform using your own data

Brought to you by

Zetaris
The Networked Data Platform

Rod Beecham
Jack Steele

About Zetaris

Zetaris was founded in 2013 to develop the next generation enterprise database and analytics platform. The Networked Data Platform, combining analytical data virtualisation and massively parallel processing, is an enabling technology that can sit anywhere — on premise, in the cloud, or as a hosted solution — while combining disparate data sources for deep analysis. By minimising data movement, the Networked Data Platform allows real-time analytics over the largest volumes of disparate data in any combination, accelerating jobs that would otherwise be long and costly, reducing time to insight, and maximising return on your IT investments while minimising cost.

For more information, visit Zetaris at: https://www.zetaris.com/.

The Networked Data Platform

Zetaris Special Edition

by Rod Beecham
Jack Steele

for dummies®
A Wiley Brand

The Networked Data Platform For Dummies®, Zetaris Special Edition

Published by: **John Wiley & Sons Australia, Ltd,** 42 McDougall Street, Milton Qld 4064

www.dummies.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION, WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANISATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANISATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

ISBN 978-0-730-39465-5 (pbk); ISBN 978-0-730-39467-9 (ebk)

Table of Contents

Publisher's Acknowledgements

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your busness or organisation, contact `info@dummies.biz` or visit `www.wiley.com/go/custompub`. For details on licensing the *For Dummies* brand for products or services, contact `BrandedRights&Licenses@Wiley.com`.

Some of the people who helped bring this book to market include the following:

Project Editor: Kerry Laundon

Acquisitions Editor: Ashley Coffey

Editorial Manager: Rev Mengle

Business Development Representative: Clare Dowdell

Production Editor: Tamilmani Varadharaj

Introduction

The promise of data-driven innovation is that it provides anyone in an organisation with access to all the relevant data they need for insight and sound decision-making.

However, to consolidate data requires careful planning and a great deal of costly time and effort, by which time new data and new data sources have appeared, making the process redundant.

Forrester analyst Noel Yuhanna introduced the term ‘data virtualisation’ in 2006 in a paper positing the concept of an information fabric architecture.

Operational Data Virtualisation, however, has its limitations. The future is the Networked Data Platform, which is based on Analytical Data Virtualisation.

The *Networked Data Platform* is a technology enabling all organisational data to be analysed in any combination. It is not a tool; instead, it is infrastructure, designed to reduce time and duplication and capable of performing deep analytics over the largest data volumes quickly.

The Networked Data Platform represents speed, scalability and platform independence. It realises the latent power of your existing IT infrastructure, enabling you to make informed decisions on all the available data in real time.

About This Book

The Networked Data Platform For Dummies contains five chapters:

- Chapter 1 explains why centralising data can’t solve the problem of managing large quantities of data.
- Chapter 2 covers how the Networked Data Platform works.
- Chapter 3 considers the many benefits of using the Networked Data Platform.
- Chapter 4 shows you how to get started with using the Networked Data Platform.
- Chapter 5 shares some ways in which you can unleash the potential of the Networked Data Platform.

Foolish Assumptions

Data management is complex. This book aims to explain key terms and make things as clear and as simple as possible. If you rely on complete, accurate and up-to-date data to do your job, this book is for you.

Icons Used in This Book

Throughout this book, we occasionally use the following special icons to draw attention to key information:

REMEMBER

This icon indicates a key feature of the Networked Data Platform that you should commit to memory.

TECHNICAL STUFF

While we aim to steer clear of jargon, this icon indicates a technical term that needs to be understood.

TIP

This icon indicates a point you may find useful as you get acquainted with the Networked Data Platform.

WARNING

This icon alerts you to something in the labyrinthine world of data you should beware of.

Beyond the Book

The Networked Data Platform For Dummies shows you how to get started with using the Networked Data Platform in Chapter 4. We encourage you to develop your expertise by continuing to try new things when you use it.

For further information about the Networked Data Platform, including white papers, use cases, podcasts and customer stories, visit `www.zetaris.com`. You can also find us on LinkedIn: `www.linkedin.com/company/zetaris`.

IN THIS CHAPTER

- » **Understanding how data platforms have evolved**
- » **Considering the challenges of managing data**
- » **Analysing why centralising data cannot solve the problem**
- » **Integrating the six data delivery styles**

Chapter 1
Introducing the Networked Data Platform

Your business needs to move quickly, making effective data-driven decisions. But you have a problem. Your business has too much data in too many different places across the network, making it hard to access — which, ultimately, slows you down. The old method of moving all data into a single source of truth is slow and costly, and it puts your data security at risk.

Imagine having a single view of your data across your entire business, no matter where the data is stored, enabling instant analytics. With the Networked Data Platform, you can.

In this chapter, you explore how and why data has become fundamental to enterprise decision-making, and how and why the data explosion has made effective enterprise decision-making challenging. You discover why attempting to centralise all data is not the answer, the different ways in which data can be integrated, and how the Networked Data Platform optimises these ways of integration.

A Brief History of Data Analytics Platforms

Enterprises are organisations, whether public, private or not-for-profit, that exist to achieve strategic goals. To survive and thrive, enterprises need to make decisions. The basis of those decisions is information. Information derives from *data* (digitally stored records of customers, clients, transactions and the like), and data nowadays is located almost exclusively in computer systems and associated digital devices.

The first dedicated decision-support systems emerged for financial planning, airline scheduling and logistics. Effective flight-scheduling, for example, depends on a variety of complex factors — who travels where, by what route, on what day(s) and at what time of day, as well as local weather conditions, time of year, fleet availability, re-fuelling facilities, and so on. Consolidating and analysing all the available data relating to these factors allows an airline to schedule the right flights at the right times to maximise revenue. Employing data in this way, to derive actionable insights, is known as *Business Intelligence* (BI).

But suppose the data is wrong? If the fleet availability data is not up to date, there may not be an aircraft available to make the flight. If the weather information is not up to date, a flight may encounter dangerous conditions. If the name of the re-fuelling facility is incorrect, an aircraft may land that can't be re-fuelled.

WARNING

BI is only as good as the data from which it is derived, and the timeliness of that data.

As the technology developed, a number of general-purpose, off-the-shelf BI products appeared. However, only high-level managers and specialists were able to access those products. Accordingly, subsequent generations of BI tools have focused on ease of use and user-friendliness. Many such products are readily available today; however, these are typically front-end products that can access only some data sources, not all. They may be employed in parts of the enterprise, but they don't allow a view of the whole enterprise.

The Promise of Data-driven Innovation

Data is the blueprint of innovation. No longer a passive entity filling up the archives, it is the necessary basis of enterprise decision-making.

Sending emails, making telephone calls, collecting information for campaigns . . . each day, people create a massive amount of data just by going about their business. The advent of the internet has generated exponential growth in the global computer-using community — computer users have ever-expanding access to computing power and bandwidth. The interaction of these users with internet applications results in unprecedented quantities of data and transactions.

This data explosion continues to accelerate. The core social networks — Facebook, Twitter and LinkedIn — have created new channels through which people can communicate and interact, resulting in correspondingly large datasets and transaction volumes. Specialised social networks are also on the rise, offering everything from match-making sites to 'buy-sell' applications that have the potential to generate their own micro-economies.

Advanced mobile devices have contributed to this data explosion, too. Smartphones and tablets generate vast streams of data, transactions, application interactions and messages, and they're popular around the world. This is the era of Big Data.

Big Data is large, complex datasets, especially from new data sources. These datasets are so large that traditional data processing software can't manage them.

REMEMBER

Big Data represents a great opportunity for enterprises. The promise of data-driven innovation is that it provides anyone in the enterprise with access to all the relevant data for insight and sound decision-making. You have the potential to understand not just what your most profitable business lines are but *why* they are the most profitable. You have the potential to understand not just where your competitors are doing well but *why* they are doing well in those areas. You can determine that the behaviour of your customers is changing and, more importantly, *why* it is changing. These are key insights for continuing success and competitive advantage.

Data-driven innovation is the exploitation of Big Data to meet a measurable need. An example comes from the oil giant, Shell. Information about revenues from Shell's fuel and non-fuel businesses revealed that 20 per cent of its products were delivering 80 per cent of its sales, enabling Shell to make significant improvements to its margin, turnover, deals with suppliers and product master file management, which helped reduce its working capital.

Enterprises have long been aware of the benefits that can accrue from the marshalling and analysis of Big Data. Industries have grown around data consolidation in the endeavour to realise those benefits, both within enterprises and within the information technology sector itself, where firms providing infrastructure, data centres and Data-as-a-Service (DaaS) have proliferated.

However, as has so often been the case in the history of information technology, marshalling and analysing Big Data is like trying to change the tyres on a moving car — neither data nor the market stand still.

Why a 'Centralise-only' Approach Has Failed

Data is not uniform. It comes in many different formats. This is one of the reasons why BI tools are employed to analyse sub-sets of data. The tools can't connect to everything, so they're used specifically to analyse financial data, marketing data or operational data. Consolidating these discrete sets of data requires careful planning and a great deal of costly time and effort.

Adding another layer of complexity is the fact that, even within a particular area of an enterprise, you can find different data types (and not all of them can be accessed using a BI tool):

- **Structured data** adheres to a pre-defined model. It is stored in a tabular format, with relationships between the different rows and columns. Common examples of structured data include Excel files and structured query language (SQL) databases. Each of these have structured rows and columns that can be sorted. The earliest database management systems (DBMSs) stored, accessed and processed structured data.

- **Unstructured data** doesn't adhere to a pre-defined model and/or is not organised in a pre-defined manner. It is typically text-heavy but may contain numerical data. Common examples of unstructured data include audio files, video files or NoSQL databases.

  REMEMBER

  The ability to extract value from unstructured data is one of the main drivers behind the rapid growth of Big Data.
- **Semi-structured data** is a form of structured data that doesn't conform to the formal structure of the data models associated with relational databases or other forms of data tables but contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. JavaScript Object Notation (JSON), Extensible Markup Language (XML) and comma-separated values (CSV) files are forms of semi-structured data.
- **Ad hoc data** is based on application programming interfaces (APIs). An *API* is a mechanism allowing interactions between two applications. Web services, for example, may use Representational State Transfer (REST) or Simple Object Access Protocol (SOAP) to provide access. Many Software-as-a-Service (SaaS) solutions provide data via APIs only.

Bringing these different types of data together in a form in which they can be accessed and analysed by a BI tool is called *data centralisation*. To centralise Big Data, you must extract it from its source, transform it into a pre-determined format and load it into a centralised repository. This process is known as extract-transform-load (ETL). For large enterprises sitting on huge volumes of data, such an undertaking is enormously complex, error-prone and costly. Even if successful, masses of new data will have been acquired before you can expect to complete your centralisation project. To complicate matters further, the largest enterprises may have not one but many different data warehouses, which exposes the futility of data centralisation. To centralise Big Data continuously is to commit to an endless and unwinnable game of catch-up.

WARNING

Extracting, transforming and loading all your data into a centralised repository (ETL) is complex, expensive, high-risk and, ultimately, futile. New data is being created all the time.

One approach that may appear to offer a solution is cloud computing. *Cloud computing* is the delivery of on-demand computing services, from applications to storage and processing power, typically provided over the internet and on a pay-as-you-go basis. However, even though cloud computing reduces expenditure on IT hardware and facilities, provides flexible payment options, and facilitates enterprise-wide mobility and collaboration, it can't solve your data problems. Ultimately, the cloud is just another place to put your data.

REMEMBER

Data centralisation has its place, but centralisation can't, by itself, meet the challenges presented by Big Data. This is in part because different types of data require different methods of integration.

The Six Data Delivery Styles

Data centralisation is the process of putting the data in one place. *Data integration* is the process of combining the centralised data so that it can be viewed as a whole. It can then be delivered in six ways:

1. **Bulk/batch data movement** acts as a support mechanism for ETL processes to consolidate data from primary databases. It involves bulk and/or batch data extraction that draws data from across system and organisational data stores. A simple example is your electricity bill. Your electrical consumption data is collected over a set period of time before being processed as a batch in the form of your bill.
2. **Data replication/synchronisation** is the frequent copying of data from one database to another to allow all users to share the same level of information, resulting in a distributed database that enables user access to the data relevant to their own tasks. This provides data synchronisation, which enables users to manage growing data volumes while gaining access to real-time information. An example is when you call your electricity provider to confirm that your address has changed. The customer service person may update your address on a system, but your address only updates across all the company's systems if effective data synchronisation is in place.

3. **Message-oriented data movement** groups data into messages that applications can read so data can be exchanged in real time. In effect, message-oriented data movement works like a phone call — if you ring someone but they're not there, you leave a message for them, during which you can share the reason you called. Similarly, one computing device sends information to another and, although the receiving device may not act on the information immediately, the information is not lost.
4. **ESB data movement** refers to the employment of an *enterprise service bus* (ESB) — which acts as a sort of telephone switchboard between applications — in service-oriented architecture (SOA). *SOA* defines a way to make software components reusable through service interfaces. The old protocols and proprietary data formats of legacy systems and systems of record are translated and integrated on the fly by the ESB to work with the SOA. If, for example, an enterprise has two different billing systems that need to be connected, an ESB can enable the two systems to communicate through the service layer, saving the enterprise both money (no coding required) and time.
5. **Data virtualisation** allows users to create a virtual abstract layer that can be mirrored to provide one single view of all the data that resides in a database, instead of having to run through the process of ETL to load the data into an analytical framework. This is of great benefit to complex enterprises with many disparate data sources, such as a stock exchange, where accurate, up-to-the-minute trading data is essential. With no time to centralise or integrate the data physically, the exchange can use data virtualisation to ensure speed and accuracy.
6. **Stream data integration** refers to the implementation of a data pipeline to ingest, filter, transform, enrich and then store the data in a target database or file to be analysed later. The need for stream data integration has emerged due to the increase in information sources that didn't exist a decade or two ago, such as mobile devices, Internet of Things (IoT) sensors and social media.

As these different data delivery styles indicate, data centralisation and data de-centralisation are not a case of 'either/or'. Indeed, one copy of data away from its source is usually necessary — but continuous copying is not. Continuous copying, however, is commonly the case, with transactional data, such as records of product purchases, being copied to an operational data store, then copied to a data warehouse, then copied to a data mart, and then copied to a CSV file for a business report.

The Networked Data Platform

The *Networked Data Platform* is an enterprise-wide harmonisation of the six data delivery styles described in the preceding section. It works not through the data itself, but through metadata.

Metadata is data about data. For example, technical metadata around a customer identification number may include the table in which the number resides and the type of data it is (alphabetical, numerical, and so on). The source and destination of an email, the servers the email touched, and the date and time the email was sent represent metadata — as distinct from the data, which is the content of the message itself.

REMEMBER

Metadata is the key to data virtualisation. The fundamental point of data virtualisation is that the data remains in its source location — it does not have to be copied or moved.

TECHNICAL STUFF

Data virtualisation operates by means of a federated metadata layer. To understand what a federated metadata layer is you need to know what a database schema is. A *database schema* represents the logical configuration of all or part of a relational database. It can exist both as a visual representation and as a set of formulas, known as integrity constraints, which govern a database. These formulas are expressed in a data definition language, such as SQL. A database schema indicates how the entities that make up the database relate to one another, including tables, views and stored procedures. The schema is a sub-set of metadata.

A *federated metadata layer* is simply an aggregation of many database schemas to enable the different data sources to be accessed as a whole. The data itself is not moved or changed in any way but, by means of the federated metadata layer, it can be accessed and analysed as if the different data types in different locations have been consolidated in a single format.

The Networked Data Platform makes extensive use of data virtualisation, which is enormously helpful in meeting the challenge of Big Data. However, data virtualisation itself is a nuanced concept.

Operational Data Virtualisation (ODV), as the name implies, is focused on operational database functions (for example, INSERT, UPDATE, DELETE) rather than analytical functions. It reduces the overhead associated with task execution (the process of a client sending the task, the server processing the task, and the server returning the response to the client) by executing more than one task at a time.

However, ODV has its limitations. Data governance cannot be applied universally or consistently. The *BI cube* (the analytical range of the BI tool) can be expanded, but only as far as the virtualisation software allows. Data pipelining remains largely physical, so ODV can only be applied to structured data.

Analytical Data Virtualisation (ADV), by contrast, employs *massively parallel processing* (MPP), which allows very large volumes of data to be interrogated by a single query, avoiding multiple trips back and forth between the application server and the database server. More importantly, it can process different types of data — structured, unstructured, semi-structured and ad hoc. ADV breaks down large tasks into smaller, independent parts that can be executed simultaneously by multiple processors communicating through a shared memory. This increases available computing power for faster processing. ADV also exploits *dynamic partitioning* — a technique by which multiple processes can be run simultaneously by a single central processing unit (CPU) — to increase processing speed and maximise the efficiency of available computing resources.

REMEMBER

ODV works with structured data only, while ADV works across different data types.

The Networked Data Platform is biased towards ADV. It can ingest and store the data itself where necessary (historical data, for example), but its key differentiator is its enablement of analytical workloads over the biggest of Big Data (meaning federated data in many places). It enables a bank, for example, to trace the hidden patterns running through trillions of individual transactions.

ADV achieves this by means of:

- **Federated query optimisation,** which enables an enterprise to analyse its interactions with an individual customer across all its data sources, using its computing resources most efficiently
- **AI-automated and rules-based data governance,** which ensures that no query inadvertently violates the customer's privacy
- **Self-service analytics,** which means that any employee of the enterprise, whatever their level of expertise with computers, can undertake deep analysis of the customer's activities
- **Virtual data pipelining,** which allows data from disparate sources to be analysed with filtering and features, which ensures that the data is accurate and the queries do not fail

Imagine a bank (call it Big Bank) has an individual customer, Angela, who has a mortgage, a cheque account, a savings account and a line of credit. A review of Angela's spending patterns, income, savings balance, available credit, loans, credit score, level of risk and related activity on social media reveals that she loves to cook, enjoys visiting gourmet restaurants, posts online about her dining experiences and would like to own a restaurant-style six-ring gas stove. Analytics reveal that Angela has recently made a number of household item purchases because she is renovating her kitchen. Big Bank, knowing Angela is a responsible person with a steady job and a good credit history, can now offer to extend Angela's line of credit to enable her to buy the stove. It can also share the information about Angela's financial history and preferences with Angela herself to help her determine her financial position, saving her from the labour-intensive process of retrieving receipts and payslips. Thanks to ADV, Angela can bring up a comprehensive view of her financial status at the touch of a button on her Big Bank app.

IN THIS CHAPTER

- » **Introducing the ADV layer**
- » **Taking your query to the data**
- » **Accelerating time to value**

Chapter 2
Understanding How the Networked Data Platform Works

The Networked Data Platform in action enables an enterprise to perform deep analytics on its data by joining and querying all data types from any source in any combination.

In this chapter, you find out how Analytical Data Virtualisation (ADV) enables you to query and analyse your data sources wherever they are and eliminates the need to centralise all your data. You discover how this serves to reduce cost and accelerate time to value. First, however, we explain what we mean by the ADV layer and how this underpins the way the Networked Data Platform works.

Getting to Grips with the ADV Layer

The *ADV layer* is a bridge across data warehouses, data marts and data lakes that substitutes for an integrated physical data platform. It leverages the data sources — including, most importantly, all data types — to maximise the availability and exploitation of enterprise data.

ADV resembles views in the world of relational database management systems (RDBMSs). In an RDBMS, rows and columns from different tables can be joined to create a virtual table. ADV acts similarly, but across all data sources — the views are not limited to data resident at a single source or to data of a single type. When an application queries the ADV layer, metadata is pulled from the various source systems and the results are transmitted without any duplication of data to the ADV layer. For example, consider a service enterprise, such as a telecommunications provider. Customers encountering problems call the customer support desk, complete an online form or send an email, and the support personnel then record the information received on a pre-defined form (a problem ticket) for prioritisation and action by technical personnel. The process is time-consuming, labour-intensive and inefficient. An ADV layer, however, can enable the three sources of information — voice, online form and email — to be reviewed and prioritised in real-time using machine learning (ML), saving time and eliminating human error.

Beyond the ADV layer, the benefits reside in powerful features such as joining heterogeneous systems (for example, the three mentioned in the example) and handling massive complex transformations.

REMEMBER

The ADV layer enriches analytical data solutions by bringing agility to the development work, parsing the semi-structured datasets and integrating with third-party systems.

ADV can be logically positioned in any or all of three places in a typical data analytics architecture (see Figure 2-1):

- **Connectivity layer:** Connects to all the sources in a few clicks and creates database objects that can be consumed by downstream systems and processes.
- **Semantic layer:** Holds to all the sources in a few clicks and creates database objects that can be consumed by downstream systems and enforcement.
- **Data layer:** Supplements the existing physical data store (or *enterprise data warehouse*) with additional information from third-party or business-owned datasets.

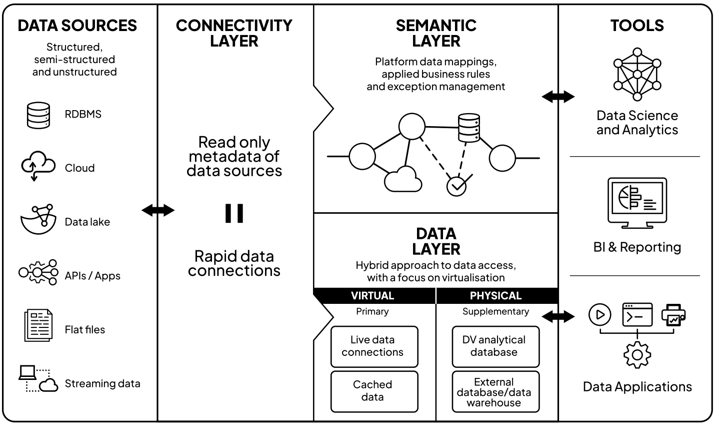

DATA SOURCES
Structured, semi-structured and unstructured
RDBMS
Cloud
Data lake
APIs / Apps
Flat files
Streaming data
CONNECTIVITY LAYER
Read only metadata of data sources
Rapid data connections
SEMANTIC LAYER
Platform data mappings, applied business rules and exception management
DATA LAYER
Hybrid approach to data access, with a focus on virtualisation
VIRTUAL
Primary
Live data connections
Cached data
PHYSICAL
Supplementary
DV analytical database
External database/data warehouse
TOOLS
Data Science and Analytics
BI & Reporting
Data Applications

FIGURE 2-1: Positioning the ADV layer.

Taking Your Query to the Data

A *query* is a request for data or information from a database table or combination of tables. Traditionally, the data must be physically present in those tables to be interrogated.

Figure 2-2 shows the continuous copying and transformation of data discussed in Chapter 1. The source data is copied to a staging area, then to a data warehouse, then to a data mart (or some other data repository) before it is ready for analysis.

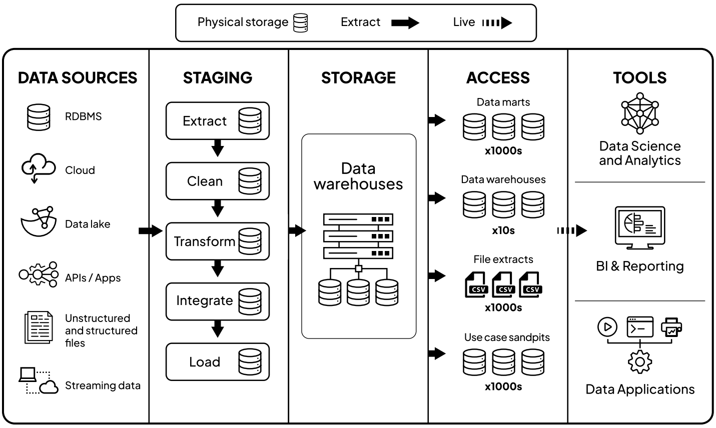


FIGURE 2-2: Taking your data to the query.

The Networked Data Platform, by contrast, enables you to take your query to the data, eliminating these duplications (see Figure 2-3).

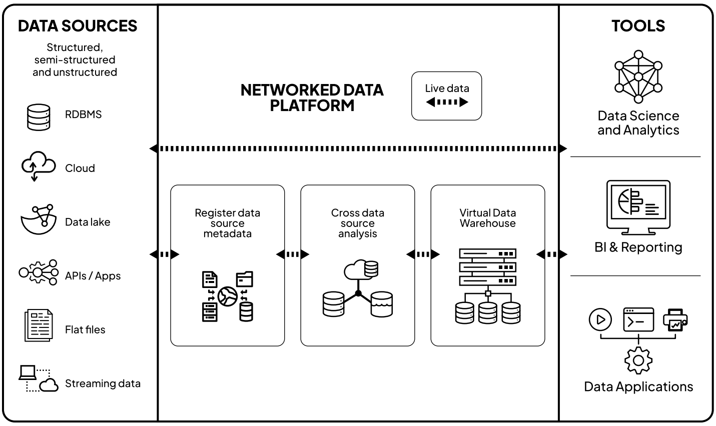


FIGURE 2-3: Taking your query to the data.

The ADV layer of the Networked Data Platform allows you to join disparate data sources in the way we have seen — through virtualisation over metadata. The data stays where it is but can be queried and analysed as if it were consolidated at a single source.

How is that possible?

The Networked Data Platform connects to the source systems via database connectors — typically, JDBC or ODBC — and your enterprise's Business Intelligence (BI) tools connect to the Networked Data Platform. No data extraction, no data transformation and no copying of data to another repository is required.

TECHNICAL STUFF

JDBC stands for Java Database Connectivity. A *JDBC driver* is a software component enabling a Java application to interact with a database. ODBC stands for Open Database Connectivity. An *ODBC driver* allows applications to access data in database management systems using structured query language (SQL) as a standard for accessing the data.

Your BI tool has native connectors enabling it to link directly to various file types and server types, so you can access your data through the Networked Data Platform using your existing BI tool. Because the Networked Data Platform virtually combines all your enterprise data, you need only establish one connection — between your BI tool and the Networked Data Platform — to access, join and analyse all your data assets in real time. Anything your BI tool could not connect to before is now available via the Networked Data Platform.

The Networked Data Platform ingests the metadata from the source systems to generate a virtual schema. A *virtual schema* is a dynamically created logical configuration of all the entities across disparate data sources, enabling them to be manipulated as if they were physically consolidated. The result is a Virtual Data Mart (VDM).

REMEMBER

A *VDM* is a group of physically separate data sources that can be analysed as if they were a single, consolidated data source.

Figure 2-4 shows how the Networked Data Platform enables disparate data sources to be queried in real time. The data sources are registered with the Networked Data Platform, enabling virtual schemas to be created from source metadata. Users can then query the data sources in any combination as if they were physically consolidated.

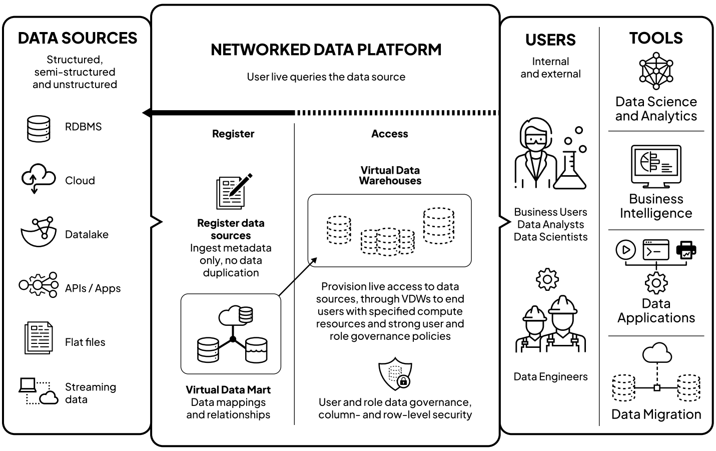

DATA SOURCES
Structured, semi-structured and unstructured
RDBMS
Cloud
Datalake
APIs / Apps
Flat files
Streaming data
NETWORKED DATA PLATFORM
User live queries the data source
Register
Register data sources
Ingest metadata only, no data duplication
Virtual Data Mart
Data mappings and relationships
Access
Virtual Data Warehouses
Provision live access to data sources, through VDWs to end users with specified compute resources and strong user and role governance policies
User and role data governance, column- and row-level security
USERS
Internal and external
Business Users
Data Analysts
Data Scientists
Data Engineers
TOOLS
Data Science and Analytics
Business Intelligence
Data Applications
Data Migration

FIGURE 2-4: The Networked Data Platform.

You see how this stops the game of catch-up? You don't need to keep copying fresh data to a centralised repository. By not moving all the data, the Networked Data Platform eliminates complex data integration, allowing you to join and query the most up-to-date data from all your data sources simultaneously.

Key Features of the Networked Data Platform

The Networked Data Platform is deep technology. It's the solution to the problems created for enterprises by the data explosion and the proliferation of IT products — databases, network designs, architectural models, applications — that have evolved to deal with them.

Four major features of the Networked Data Platform should be emphasised:

- **Flexibility:** You don't have to move or copy data, but you can if necessary.
- **Data integrity:** The data is not compromised by movement and transformation.
- **Analytical speed:** Queries can be optimised across disparate data sources and data types.
- **Data governance:** Security, access controls and business rules can be enforced at the level of the individual query.

Move data when you need to

Some data does need to be copied. Historical data, for example, which is obviously important when undertaking deep analytics, is usually offloaded from source systems so they hold only a couple of months' worth of the most recent data. The Networked Data Platform can ingest and store such data in a massively parallel processing (MPP) database so that it is readily available for analysis. The MPP database is optimised for analytical workloads, aggregating and processing very large datasets. In this way, data not readily available at source can be accessed and queried within the Networked Data Platform in real time, either by itself or in combination with other data sources.

TECHNICAL STUFF

MPP is a storage structure designed to handle the co-ordinated processing of program operations by multiple processors. This co-ordinated processing can work on different parts of a program, with each processor using its own operating system and memory. This allows MPP databases to handle massive amounts of data and provide much faster analytics based on large datasets.

Preserve data integrity by analysing the data at source

Data exists in multiple formats (and many vendors store data in proprietary formats, effectively chaining you to a single solution). Some siloed or on-premises data is attached to older legacy systems that cannot be moved without re-engineering those systems.

When data is stored in different systems, interoperability with outside systems and BI tools may be limited. Security and entitlements are difficult to maintain when you're merging data from many silos that may have different users and configurations.

By analysing the data at source, the Networked Data Platform integrates and conforms with the data management, governance and security practices that your enterprise IT team have meticulously built for legacy on-premises solutions. If you don't have to physically move your data, you can maintain data integrity and eliminate the risks associated with physical data movement, as well as dramatically reduce your data infrastructure running costs, improve your query performance and significantly deepen insights across the organisation. Chapter 3 talks about these benefits in more detail.

Run large, complex analytics at speed

In data analytics, *scaling* refers not simply to how well a solution can handle growing volumes of data, but also to how well it can handle large and complex query volumes, more users, a greater variety of queries, and a wider variety of ML models. Traditional database management systems (DBMSs) employ an automated process known as *query optimisation* to enable optimal performance.

The Networked Data Platform, however, employs an *Analytical Query Optimiser* (see Chapter 3) that analyses modes of access, data structures and orders of processing across all the data sources required by the query (rather than just a single data source across a single network and system). With the Networked Data Platform, you can optimise speed and scalability for whatever query you are running across any combination of data sources.

Enforce comprehensive data governance

Data governance is a set of principles and practices that ensures high quality through the complete lifecycle of an enterprise's data. The Networked Data Platform not only inherits all client security and access controls but implements policy-based governance (PBG) and query behaviour control, where policy rules are implemented in a policy register that can be automatically referenced for compliance by every query, algorithm or code operation. *PBG* ensures that business rules are enforced in tandem with regular security and access controls. For example, a user may have access rights to a number of data sources but attempt to run a query that violates a business rule (say, a query seeking to determine whether a relationship exists between customers' buying behaviour and their sexual orientation, which would be a breach of privacy). PBG ensures that any such query isn't executed.

IN THIS CHAPTER

- Maintaining your existing IT infrastructure
- Saving your business time and money
- Adapting to an ever-changing data environment
- Anticipating and preventing security breaches

Chapter 3 Exploring the Benefits of the Networked Data Platform

Information technology assets are powerful, but also expensive. As anyone managing a large enterprise knows, *total cost of ownership* (TCO), meaning the purchase price of an asset plus the costs of operation, is a very important calculation. The business benefits provided by the asset must outweigh the TCO. If not, the asset becomes a drag on the bottom line.

In this chapter, you discover the many benefits of the Networked Data Platform, including how it can help to drive down your IT costs and maximise the return on your IT investments.

Deploy without Replacing Your Existing Data Infrastructure

Large enterprises, in particular, have very deep and extensive investments in IT infrastructure, including multiple data warehouses, Hadoop distributions, Software-as-a-Service (SaaS)/ cloud applications, and messaging, flat-file and legacy systems.

In effect, the Networked Data Platform is an analytical engine powered by your existing data infrastructure. It connects to all the disparate data sources, ingesting metadata to create a virtual schema (refer to Chapter 2 for more on this), reinforcing data security, and optimising and accelerating query performance. The semantic layer of the Networked Data Platform, in turn, allows all your data consumers to manipulate and interrogate your total data assets in real time.

REMEMBER

The Networked Data Platform is not a replacement system. It is an enabling system, allowing you to realise the full value of all your IT investments.

By enabling a simultaneous view of all data sources through Analytical Data Virtualisation (ADV), the Networked Data Platform eliminates all the headaches associated with a traditional data warehousing effort. The Networked Data Platform allows users a complete view of all data, wherever it is, via a single web interface and the user's preferred tools. Historical data can be kept in flat-file storage, low-cost databases or within the Networked Data Platform's massively parallel processing (MPP) database (refer to Chapter 2 for more on MPP).

Any outside system or Business Intelligence (BI) tool can be used to examine the data because the Networked Data Platform feeds whatever tool is being used at the front end. It employs common connectivity standards, application programming interfaces (APIs) and network connectivity while applying security rules at the level of the individual user.

You can also structure existing and new data assets according to specific business consumption requirements. This avoids the need to re-design your data architecture according to the configuration of a particular solution. In other words, you have no vendor lock-in.

The Networked Data Platform's data publishing capability automatically works out the optimal way of presenting the virtual data views to the BI tool employed. It is use-case specific.

REMEMBER

By connecting to existing data sources without moving, ingesting or duplicating data, the Networked Data Platform eliminates data redundancy to provide a single, unified view. This represents massive cost savings in terms of data preparation, infrastructure, licensing and support.

Handle Workloads with the Analytical Query Optimiser

In IT, a *workload* is all the resources and processes necessary to make an application useful. When you access large datasets, querying often involves mixed workloads, which can be simultaneous data loading and end-user querying, a combination of transaction processing and intensive analytical querying, or the joining of a row or column from a small data source to massive data warehouse tables.

Database administrators (DBAs) strive constantly to manage the performance of these mixed workloads, which requires considerable expertise. DBAs need to know both the users and the applications running the workloads; whether the workloads are batch, ad hoc and/or resource intensive; when the workloads are being run; the business priority of each workload; and the sources of any performance issues.

The Networked Data Platform's Analytical Query Optimiser meets the challenges of managing mixed workloads automatically. The Analytical Query Optimiser calculates query costs, applies business rules, and employs heuristics and artificial intelligence (AI) to handle mixed workloads and complex data joins across a distributed computing environment.

TECHNICAL STUFF

In computing, a *heuristic* is a technique designed for solving a problem quickly when traditional methods are too slow, or for finding an approximate solution when traditional methods fail to find an exact solution.

The Analytical Query Optimiser simplifies operations and facilitates workload consolidation. By consolidating mixed workloads, including critical enterprise applications and databases, the Analytical Query Optimiser drives down capital and operating costs, freeing up budget for innovation.

Reduce Data Egress with the Intelligent Adaptive Cache

The process of running a query against cloud databases can become extremely expensive as more users connect to the system and as query complexity increases. Data egress fees are one of the cloud's biggest hidden costs.

WARNING

Data egress refers to data moving out of the cloud environment. Most cloud providers allow you to input your data (ingress) at no cost, but they charge large network fees to move your data out of their environment (egress).

For example, if you're transferring a terabyte of data from your cloud provider's environment to somewhere else, you can be looking at a cost of $100 per transfer. That may not sound like much, but if you're performing the transfer ten times a day your monthly bill may be in excess of $20,000. (If your data is moving overseas, the cost can be even higher.)

The Networked Data Platform minimises these costs not only through the employment of ADV, which minimises data movement (refer to Chapter 2 for more on this), but also through the Intelligent Adaptive Cache.

TECHNICAL STUFF

In computing, a *cache* is a high-speed data storage layer that stores a sub-set of data, typically transient in nature, so that future requests for that data are served up faster than is possible by accessing the data's primary storage location. Caching allows you to reuse previously retrieved or computed data efficiently.

The *Intelligent Adaptive Cache* analyses your cloud workload and automatically implements optimisations, including intelligent local caching of commonly accessed data. Frequently used queries are cached to improve data retrieval performance by reducing the need to access the slower underlying storage layer.

You can also cache dimension tables into main memory to remove duplication and improve *JOIN performance* (the speed at which different data sources are combined) and cache fact tables to improve query performance. The Intelligent Adaptive Cache determines dimension and fact tables automatically from the query pattern to reduce query time and implicitly cache them.

TECHNICAL STUFF

A *fact table* is a primary table in a dimensional data model that stores quantitative information, or facts, about a business process. A *dimension table* stores the details of one of the facts. For example, if the fact is time, the dimension table contains details such as year, month and day.

Apply Queries in Place with Virtual Data Pipelining

A *data pipeline* is a series of processing steps enabling data to move from, say, an application to a data warehouse or from a data lake to an analytics database. It is a physical process, moving data from A to B.

The Networked Data Platform employs *Virtual Data Pipelining*, which removes the need for extract-transform-load (ETL) processing and physical data pipelines (refer to Chapter 1). Virtual Data Pipelining enables business and transformational rules to be applied to the source data in place (the data itself does not move) to ensure that the end user has a high-integrity view of the data required. Any necessary filtering, renaming and information standards are applied to the source data.

As Virtual Data Pipelines are developed by the user community, they can be joined in future data projects, reducing time to value and solving data pipelining requirements without coding.

TIP

If your enterprise builds applications with small code bases for specific purposes, you will be moving data between more and more applications, making the efficiency of your data pipelines a critical consideration in planning and development. Virtual Data Pipelining solves the problem at a stroke!

TECHNICAL STUFF

The information management and business rules applied in Virtual Data Pipelining reside in the Networked Data Platform's *Rules-based Data Quality Engine*, which ensures a high-integrity view of the data.

Physical data pipelines, whether built in-house or by an external provider, cost tens of thousands of dollars. Virtual Data Pipelining eliminates this cost.

Empower Your Users to Work with Your Data

The Networked Data Platform empowers not just the specialist data scientist and researcher but any member of the workforce capable of constructing a simple structured query language (SQL) query or even building an Excel formula.

Users can create sophisticated models (including control logic) to generate what acts in effect as an API on the information. The API can then, for example, be used by an Excel spreadsheet, a fully fledged BI suite or a visualisation tool, or be addressed by any layer of code, bespoke construction or existing application.

The Networked Data Platform also enables machine learning through algorithms that identify patterns and features in massive amounts of data, which helps you to make decisions and predictions based on new data. The Networked Data Platform minimises the effects of change by re-directing *pointers* (the database's method of locating specific records), which removes the need to overhaul the data management systems themselves. This agility means that the Networked Data Platform can adapt to change faster than any physicalised structure.

Figure 3-1 shows how the Networked Data Platform puts the data consumer at the centre. You don't need to be knowledgeable about the design and structure of databases. You don't need to be an expert on the different data types. You don't need to understand systems architecture or back-end IT infrastructure. All you need is a computer connected to the Networked Data Platform so that you can analyse all your enterprise data in any combination you wish.

REMEMBER

The Networked Data Platform puts data consumers at the centre, which democratises enterprise data. Any user can perform deep analytics.

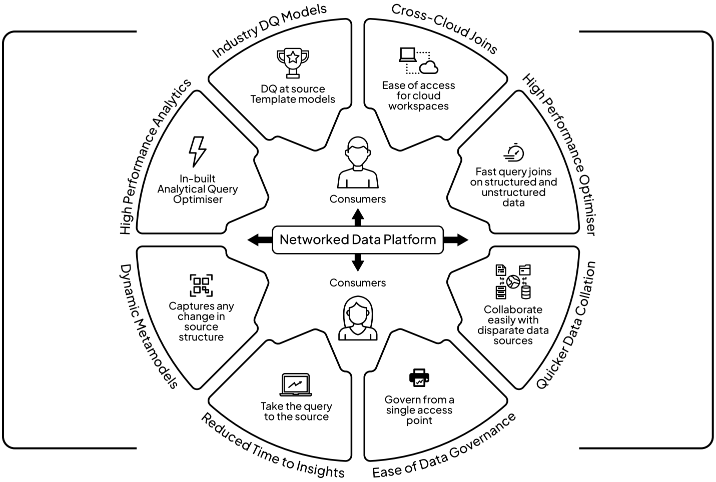

Industry DQ Models
DQ at source Template models
Cross-Cloud Joins
Ease of access for cloud workspaces
High Performance Analytics
In-built Analytical Query Optimiser
Consumers
High Performance Optimiser
Fast query joins on structured and unstructured data
Networked Data Platform
Consumers
Dynamic Metamodels
Captures any change in source structure
Quicker Data Collation
Collaborate easily with disparate data sources
Take the query to the source
Govern from a single access point
Reduced Time to Insights
Ease of Data Governance

FIGURE 3-1: The Networked Data Platform enables any data consumer to perform deep analytics.

Stop Data Security Breaches Before They Happen

Organisations face significant regulatory and compliance risks from their data management and analytics practices. With the proliferation of data and the significant business opportunities centring around data analytics and AI, real-time policy-based governance (PBG) of data, operations and development is essential (for more on PBG, refer to Chapter 2).

The risk lies in the fact that data extracts are all over the place — on laptops, on store servers and in varying forms of encryption. Moreover, the BI tools used within the enterprise are joining data in arbitrary ways with no oversight.

The Networked Data Platform assesses every query, algorithm or data operation performed by an analyst, developer, AI agent or BI user for policy compliance before it is run. This means that the data and the operations performed on the data across your data landscape are managed in real-time.

The Networked Data Platform generates policy decisions by evaluating the query input against policies and data. For example:

- What combinations of data joins are high-risk?
- Which users can access which resources?
- Which users can perform what query or implement which algorithm?
- What AI can access which data with what operation?
- To which *subnets* (networks within the network) is data allowed to move?
- To which *clusters* (groups of two or more computers running in parallel) must a workload be deployed?
- From which *configuration files* (databases of settings for the operating system) can *binaries* (compiled code for installation) be downloaded?
- With which operating system capabilities can a *container* (a standard unit of software that packages up code and all its dependencies to allow the application to run quickly and reliably from one computing environment to another) execute?
- At what times of day can the system be accessed?

Figure 3-2 shows how the Networked Data Platform integrates the existing data policies of the enterprise with column- and row-level security based on business rules.

REMEMBER

By applying business rules at the level of the individual query, the Networked Data Platform helps prevent data security breaches before they happen.

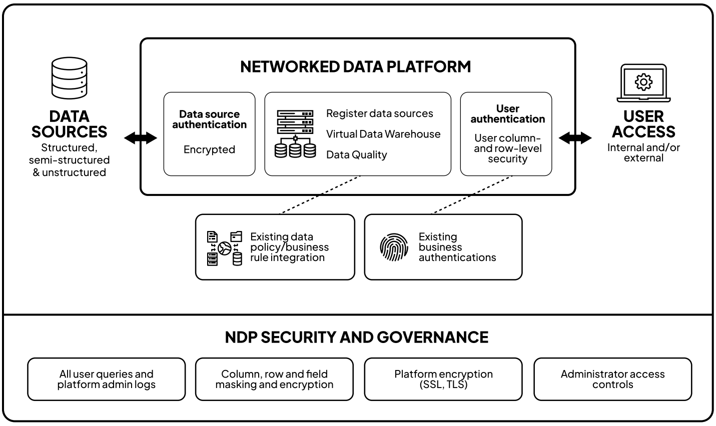

DATA SOURCES
Structured, semi-structured & unstructured
NETWORKED DATA PLATFORM
Data source authentication
Encrypted
Register data sources
Virtual Data Warehouse
Data Quality
User authentication
User column- and row-level security
USER ACCESS
Internal and/or external
Existing data policy/business rule integration
Existing business authentications
NDP SECURITY AND GOVERNANCE
All user queries and platform admin logs
Column, row and field masking and encryption
Platform encryption (SSL, TLS)
Administrator access controls

FIGURE 3-2: Data security with the Networked Data Platform.

Establish and Connect to Data Sources in Minutes, Not Months

REMEMBER

The Networked Data Platform uses virtualisation to enable data from many different sources and systems to be queried and analysed without having to consolidate the data physically in a single location.

The Networked Data Platform connects to the source systems to register the metadata of the objects and add them to its data catalogue. The data from all sources can then be interrogated using standard SQL commands.

Objects from data sources can be allocated to Virtual Data Warehouses (see Chapter 4), which can be restricted to specific users. The Virtual Data Warehouses can then be accessed to run analytics against the source systems.

Data is never stored in or ingested by the data warehouses, but the metadata for the objects that the data warehouse can access is stored.

REMEMBER

The Networked Data Platform is SaaS that can run on your preferred cloud platform. There's no installation cost — simply create a subscription and begin.

ACCELERATING DATA EXCHANGE USING THE NETWORKED DATA PLATFORM

A large enterprise saw new business opportunities in cloud data sharing. Specifically, the enterprise saw how cloud data sharing could eliminate data silos, optimise costs, streamline operations, improve customer service and turn data sharing into both a product and a product differentiator.

Data sharing has typically meant sharing *copies* of data, which creates challenges because datasets quickly become outdated and/or unsynchronised. Effective data governance becomes very difficult to enforce because both data discovery and storage costs escalate when multiple copies of data exist, and the risk of data breaches increases with every data copy made.

By eliminating the movement and copying of data, the Networked Data Platform enabled:

- Direct, real-time access to live data in a controlled environment
- The full capabilities of a physical data warehouse in a virtual environment
- Data sharing with no additional infrastructure cost
- Complete data security
- The same data to be accessed by an unlimited number of users and user groups.

The Networked Data Platform eliminated the need to move and copy data to local storage for processing, thereby de-coupling 'storage and compute' to allow the immediate availability of the latest data and dynamic scaling according to data volumes. It allowed the metadata of each data source to remain in place for dynamic ingestion, ensuring that the virtual schemas were always up to date. It also allowed unlimited concurrency (that is, multiple computations being performed at the same time).

The Networked Data Platform's virtualisation capability dramatically compressed time to value for the enterprise — as well as greatly increasing the value by removing the costs associated with moving and copying data. It ensured that the enterprise retained full control of its data while allowing external interrogation.

IN THIS CHAPTER

- **Registering with the Networked Data Platform**
- **Connecting your data**
- **Using your own BI tool**
- **Speeding up with Agile BI**

Chapter 4 Getting Started with the Networked Data Platform

The Networked Data Platform is available as Software-as-a-Service (SaaS). Enterprises can install the Networked Data Platform on their premises if they wish, but the software is also available as a hosted solution for anyone who wishes to use it. Why not give it a try now?

In this chapter, you discover how to access the Networked Data Platform, how to connect your data to it, and how to exploit it for self-service and Agile Business Intelligence (BI).

Accessing the Networked Data Platform

To deploy the Networked Data Platform, all you need is the credentials and privileges you require to access your data sources (such as access keys; a path to the files, Blob or S3 bucket; and/or database access credentials) and access to the Networked Data Platform. Data sources may include flat files, databases and streaming data, whether you are on-site or operating in the cloud. No build, no reconfiguration, no architectural redesign — the Networked Data Platform can be activated in minutes!

TECHNICAL STUFF

Azure Blob Storage is Microsoft's object storage solution for the cloud, optimised for storing massive amounts of unstructured data. An *S3 bucket* is a public cloud storage resource available in Amazon Web Service's Simple Storage Service (S3), an object storage offering. Amazon S3 buckets are similar to file folders and store objects consisting of data and its descriptive metadata.

Registering as a user of the Networked Data Platform is easy. Simply visit `https://www.zetaris.com/ndp4dummies` and click on the registration link. You will receive your access credentials via email, and then you can access the Networked Data Platform through the log-in link at `https://www.zetaris.com/ndp4dummies`.

After you log in, you see ten icons next to the Zetaris logo on the navigation bar. These icons are, from left to right:

1. NDP Fabric Builder
2. Schema Store View
3. Data Catalog
4. Query Builder
5. Virtual Data Mart
6. Virtual Data Warehouse Builder
7. Data Lineage View
8. NDP File System
9. Data Pipeline
10. User Management

You can allocate objects from data sources to virtual data warehouses, which you can restrict to specific users. You can then access the Virtual Data Warehouses and run analytics against the source systems.

REMEMBER

A *Virtual Data Warehouse* is a group of physically separate data sources that you can analyse as if it is a single consolidated data source. Data is never stored in or ingested by these data warehouses; however, the metadata for the objects that the data warehouses access *is* stored in these data warehouses. Because the data itself has not moved, and because existing access and security permissions are augmented by the Networked Data Platform's policy-based data governance, the integrity of your data is assured.

Connecting to Your Data

Before you can build your first query, you need to register your data sources so that the Networked Data Platform knows where to find them.

To register each data source and connect to it:

1. **On the navigation bar, click on the NDP Fabric Builder icon (see Figure 4-1) in the top left-hand corner of the screen. This is the first icon on the left.**

   The NDP Fabric Builder screen appears.

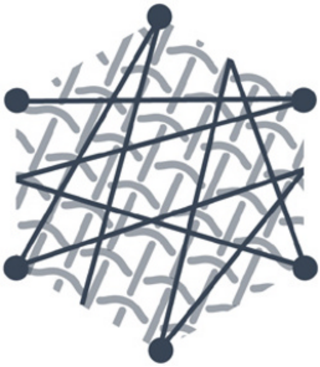


FIGURE 4-1: The NDP Fabric Builder icon.

The left-hand side of the NDP Fabric Builder screen displays sections for physical data sources, logical data sources and streaming data sources.

2. **Click on the orange plus-sign next to the type of data source you want to connect to and follow the steps in the pop-up windows that appear.**

   TIP

   When you add a logical data source, the Logical Data Sources pop-up will ask you to specify Pull Type or Push Type. Always specify Pull Type. Push Type applies only in more complex scenarios, such as linking data from an enterprise Networked Data Platform to a cloud Networked Data Platform.

   As you register each data source, it appears under the relevant section on the left-hand side of the screen.

You are now ready to create your first Virtual Data Warehouse:

1. **On the navigation bar, click on the Virtual Data Warehouse Builder icon (see Figure 4-2).**

   The Virtual Data Warehouse Builder screen appears.

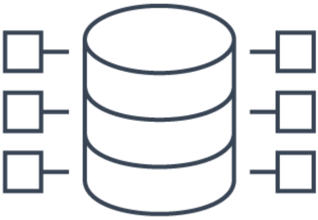


FIGURE 4-2: The Virtual Data Warehouse Builder icon.

   In the top left-hand corner of the Virtual Data Warehouse Builder screen, you see the Create button.

2. **Click on the Create button and follow the steps in the pop-ups that appear.**

   After you have created the Virtual Data Warehouse, the Status column shows it as Pending. After a few moments, Pending will change to Running, which means that the Virtual Data Warehouse is ready for interrogation.

Congratulations! You have created your first Networked Data Platform.

Adding Your BI Tool

You can access your Virtual Data Warehouse directly via the Networked Data Platform web interface or through your preferred BI tool. Your BI tool has powerful features for presenting your data as you wish — charts, graphs, histograms and the like — but, until now, it has only shown you a sub-set of your data. Once you connect to the Networked Data Platform, however, your BI tool will be able to show you everything.

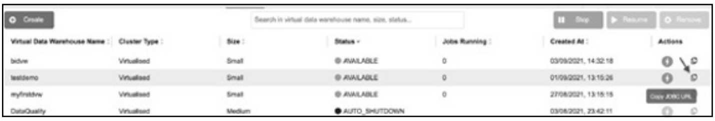
TECHNICAL STUFF

The Networked Data Platform supports both JDBC (Java Database Connectivity) and ODBC (Open Database Connectivity) drivers. A JDBC driver is a file with the extension .jar, while an ODBC driver is a file with the extension .dsn.

For the purposes of this example, imagine that your BI tool requires a JDBC driver. To connect your BI tool:

1. **Download and configure the Networked Data Platform JDBC driver, which you can access via the connection link at** `https://www.zetaris.com/ndp4dummies`**.**
2. **Follow the specific instructions for your BI tool to register the JDBC driver (the .jar file). If your BI tool requires a driver class, enter 'com.zetaris.lightning.jdbc.LightningDriver'.**

   During this process you will be provided with a JDBC URL, which you will require in Step 4.
3. **Return to the Virtual Data Warehouse Builder screen. Ensure that your Virtual Data Warehouse is Running (rather than Pending; refer to the preceding section 'Connecting to Your Data' for more on checking this) and click on 'Copy JDBC URL' (as shown in Figure 4-3).**



Virtual Data Warehouse Name	Cluster Type	Size	Status	Jobs Running	Created At	Actions
bidvw	Virtualised	Small	● AVAILABLE	0	03/09/2021, 14:32:18	
testdemo	Virtualised	Small	● AVAILABLE	0	01/09/2021, 13:15:26	
myfirstdvw	Virtualised	Small	● AVAILABLE	0	27/08/2021, 13:18:18	
DataQuality	Virtualised	Medium	● AUTO_SHUTDOWN		03/08/2021, 23:42:11	

FIGURE 4-3: Click on 'Copy JDBC URL' to connect your BI tool to your Virtual Data Warehouse.

4. **Return to your BI tool to connect to your Virtual Data Warehouse using your JDBC URL and any other required details, such as the JDBC driver name (which will be ‘com.zetaris.lightning.jdbc.LightningDriver’), your username and your password.**

Now your BI tool can be used to run queries and analytics across your data.

Enabling Agile BI

You may have heard the expression ‘Agile BI’. It refers to the application of the Agile software development model to BI projects. Essentially, *Agile BI* means taking an iterative approach to BI that yields immediate benefits, as opposed to building a solution piece by piece and realising the benefits only at the end.

The Networked Data Platform enables warp-speed Agile BI!

REMEMBER

The Networked Data Platform achieves this by eliminating the need for the data analysis and design phases of the BI project. The data stays where it is and is accessed in place, so you don't need to move your data.

The beauty of this is that if you discover you need an additional field or table in the report you are building, you don't have to start all over again.

To enable Agile BI:

1. **From the navigation bar, click on the Query Builder icon (see Figure 4-4).**

   The Query Builder screen appears.

2. **On the Query Builder screen, you can drag and drop tables from your registered data sources into the main window and create links between the fields with your cursor (see Figure 4-5).**

   The combinations of tables you create become a view that can be assigned to and accessed by anyone.

   TIP

   You can create new data combinations at the click of a button to yield the insights you want.

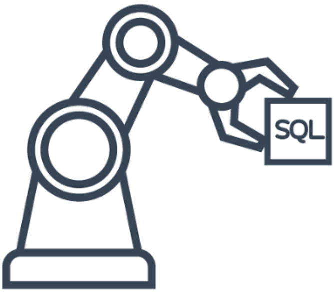
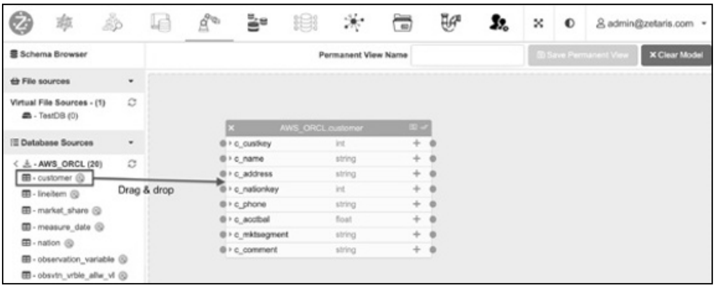


FIGURE 4-4: The Query Builder icon.



c_custkey	int
c_name	string
c_address	string
c_nationkey	int
c_phone	string
c_acctbal	float
c_mktsegment	string
c_comment	string

FIGURE 4-5: Combining your data sources for analysis.

The structured query language (SQL) you generate can be copied and saved, or it can be executed to view the result-set.

Joining disparate data types from different data sources becomes a simple matter of drag-and-drop with the Networked Data Platform.

IN THIS CHAPTER

- » Optimising query efficiency at scale
- » Assimilating data in real time
- » Surveilling your metadata

Chapter 5
Ten Ways to Unleash the Potential of Your Data

Reader, we thank you for having come this far. There has been a LOT of technical information to assimilate, but we hope you can see why this has been necessary. The Networked Data Platform is an enabling technology and, unless you understand what happens behind the scenes in your enterprise's data landscape, you may not appreciate the business value the Networked Data Platform delivers.

In this chapter, we highlight ten ways in which the Networked Data Platform can unleash the potential of your data assets.

Make Decisions Using All the Available Data

The Networked Data Platform ensures that your business decision-making is based on all the relevant data because it allows you to register all your data sources and combine them in Virtual Data Warehouses in any way you wish. You are not limited to the data your Business Intelligence (BI) tools can access directly.

Ensure the Quality of Your Data

A data quality assessment compares the actual state of a particular set of data to a desired state. The desired state is usually defined by one or more individuals or groups, standards organisations, laws and regulations, business policies, or software development policies.

Data quality assessment of all source data is a by-product of the Networked Data Platform's metadata-based approach, which ensures common data definitions linked to current and future business applications through its Rules-based Data Quality Engine (refer to Chapter 3). The Networked Data Platform includes a comprehensive exception management feature based on *root cause analysis* (identifying the underlying cause of the exception), which allows immediate identification of exceptions via a user-friendly dashboard.

Keep Your Business Running During Data Migration

Data migration is the process of moving data from one location to another, one format to another, or one application to another. Data migrations often occur when enterprises move from on-premises infrastructure and applications to cloud-based storage and applications to optimise or transform their business.

The Networked Data Platform allows 'lights-on' data migration, which means that all the data being extracted, transformed and loaded can be accessed during the migration process because it ingests metadata only, allowing the data itself to be interrogated 'in motion'.

Optimise Query Efficiency

The Networked Data Platform is built on Apache Spark, a fast and general engine for large-scale data processing. Apache Spark works with the system to distribute data across the cluster and process the data in parallel. In the Networked Data Platform it combines with the Analytical Query Optimiser to determine the most efficient execution mechanism. Every query is optimised at run time by the Networked Data Platform's Analytical Query Optimiser (refer to Chapter 3), a process involving dynamic analysis of source platform performance characteristics — which means that even the most complex queries over petabytes of data will complete within minutes.

Enforce Business Rules

Business rules are directives that define or constrain business activities. As indicated in Chapter 2, the Networked Data Platform not only integrates all client security and access controls but also implements its proprietary policy-based governance (PBG) and query behaviour controls, which enforce business rules.

For example, if a product can only be sold to people over the age of 18 in some states, PBG ensures that no one in your enterprise can inadvertently breach this condition.

TIP

Implementing the Networked Data Platform is the simplest and easiest way to ensure regulatory compliance.

Use Virtual Data Pipelining for Data Transformation

Virtual Data Pipelining (refer to Chapter 3) allows transformed versions of data to be made available in your Virtual Data Warehouse. The data target (meaning the format in which you want the data to appear) can also be made available in the Virtual Data Warehouse, as can any exceptions you generate. The Networked Data Platform enables you to simulate ETL operations in real time, so you can apply complicated business logic to your data — with the necessary transformations being made on the fly.

Assimilate Data in Real Time

Fresh data and new data sources appear all the time. The Networked Data Platform ensures that you are no longer playing catch-up. It automatically accesses the latest versions of your registered data sources, while you can add new data sources with a few clicks.

Expand the BI Cube

A BI tool, no matter how powerful or user-friendly, has limitations; for example, the tool may not integrate well with certain applications; the tool's own application programming interfaces (APIs) may be sub-optimal; or you may have to create data pipelines.

REMEMBER

The Networked Data Platform enables your BI tool to access everything, enabling user-friendly visualisations without data movement. By querying the data in place, the Networked Data Platform allows your BI tool to strut its stuff over all your data sources without having to create a patchwork of data pipelines and APIs.

Reduce Cloud Computing Costs

Cloud data warehouses can be expensive. The central processing unit (CPU) and input/output (I/O) are not always well managed across user groups and cloud domains. Duplicated and similar queries waste both time and money.

To resolve this, the Networked Data Platform identifies the data used by similar applications and creates common query caches within the Virtual Data Warehouses. By redirecting processing and data access to the Networked Data Platform, enterprises can create a 'drill-through' path to the cloud data. This drastically reduces the load on the cloud data warehouse CPU, saving you both time and money.

Enable Metadata Surveillance

Physical data warehouses and dimensional data marts are typically constructed on the basis of star schemas.

TECHNICAL STUFF

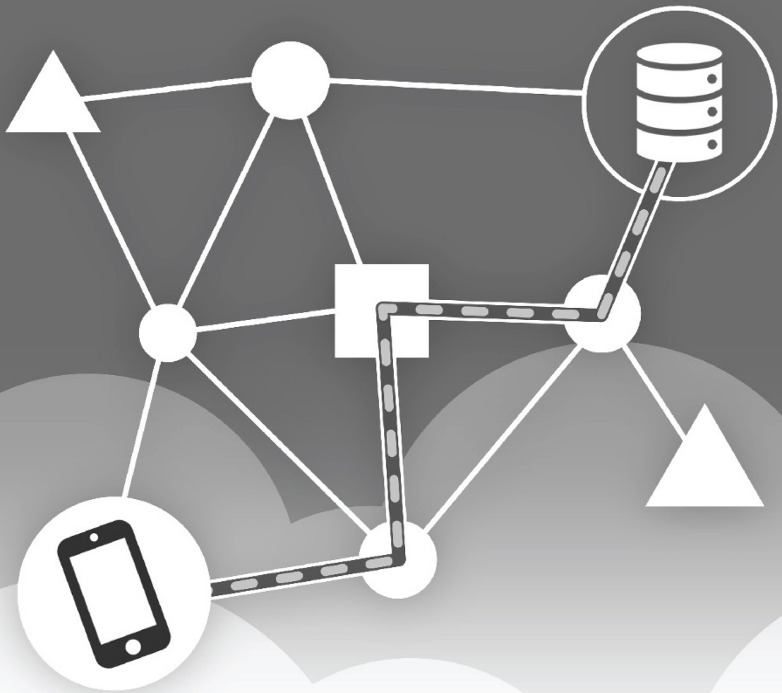
A *star schema* is a database structure optimised for use in a data warehouse that uses a single large fact table to store transactional or measured data and one or more smaller dimensional tables that store attributes about the data. It is called a star schema because the fact table sits at the centre of the logical diagram and the small dimensional tables branch off to form the points of the star.

However, the star schema is not as flexible in terms of analytical needs as a normalised data model, which reduces data redundancy and improves data integrity.

The Networked Data Platform's Schema Store tags metadata to describe all the data sources, including where they are coming from. With the Networked Data Platform, you can query and report on metadata in one logical location. The tagging itself is searchable.

REMEMBER

The Networked Data Platform enables full inspection and enrichment of your metadata through tags and descriptors, ensuring that the attributes and relations of your data are efficiently organised and their dependencies properly enforced.



Deliver reporting in a fraction of the time

Data is everywhere — in different formats, in different repositories — and new data is being generated every moment. Trying to consolidate and make sense of it all has been like playing an unwinnable game of catch-up — until now. In this book, you find the solution to this challenge — the Networked Data Platform, a ground-breaking technology that allows you to access and combine all your data, wherever it is and in whatever form, in real time for deep analytical insights.

Inside…

- Understand the problems of data centralisation
- Harmonise the six data delivery styles
- Optimise your existing IT infrastructure
- Process complex data queries in place
- Enable users to perform deep analytics
- Optimise query efficiency at scale

Zetaris
The Networked Data Platform

Rod Beecham has worked in the IT industry for 32 years. He has taught computer science and published a wide range of reviews and articles. **Jack Steele** is an experienced Data Executive. He has delivered leading data capabilities across the healthcare, banking and energy sectors.

ISBN: 978-0-730-39465-5
Not For Resale

9 780730 394655

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT