ENTITY/RELATIONSHIP PROFILING FOR SHAPING YOUR DIAMOND OF DATA PROFESSONALS

Whitepaper

SUMMARY

Responsibilities of data professionals are vast, require curiosity, perseverance, technical skill, and the ability to communicate. This is not surprising since data is regarded as the top asset of a company. Indeed, turning data into actionable insight, informed decision making and competitive advantage is the very reason why data has been so hot, so wow, so now for some years in a row.

Due to these responsibilities, data professionals are under pressure to spin straw into gold – or perhaps less allegorically – to turn raw data into value. Indeed, after examining what the actual value of data should be, and what the tasks of various data roles are in delivering this value, we will argue that current technology cannot pro-actively help with the delivery. We then describe how Entity/Relationship Profiling overcomes this limitation and can turn the ambition of data roles into reality, thereby increasing productivity, satisfying customers and employees, and ultimately making data and data professions profitable.

1. WHAT IS ENTITY/RELATIONSHIP PROFILING AND WHY IT MATTERS

The true value of data is the insight stories hidden inside them. These stories are composed of business entities, such as products or customers, and relationships between them, such as sales and acquisitions. As a consequence, every data project is ultimately about composing business entities and their relationships into a story, and every data role involved in the project must contribute towards uncovering which aspects of the entities and their relationships are important for the story. For data modelling experts this must sound familiar. Indeed, Peter Chen developed his *Entity-Relationship Model* back in the mid 1970s, the original paper continues to be the most cited paper in computer science. However, the task that data professionals face today is to extract Entity-Relationship models from data. We call this task: ***Entity-Relationship Profiling***. At its core are the two problems of computing 1) all the different ways every business entity can be identified uniquely in a data set, and 2) all the different ways column combinations reference one another across different data sets. For example, we may want to know all the different ways in which we can uniquely identify customers (Problem 1), and all the different ways in which customers appear across different data sets (Problem 2). Indeed, without such knowledge one could hardly claim to “Know your Customer”, or – more generally – “Know your Data”.

Copyright © DataViadotto Limited

Unfortunately, current data profiling technology falls short in delivering these tasks. While relationship profiling constitutes one of the three types of data profiling, current tools only offer limited support in the form of relationship analysis. This means the user must provide a relationship they are interested in, and the tool then analyses the degree by which the relationship actually holds on the data. This is merely reactive to a user-specified input, and can be accomplished with simple SQL queries. In a previous white paper, we argued why *DataViadotto Profiler* is the only true Entity/Relationship profiler, distinct by its ability of performing **D**ata-driven, **E**xpressive, **S**ample-based, **M**etric, **I**ndustry-compliant **D**iscovery at **S**cale. This is achieved by making available the latest academic advances on the underlying computational problems.

We now outline different data roles, their main tasks and contexts within an organization. Depending on the organization, one person may take on several of these roles, or some of the roles may not be necessary or have different focus points. The roles are summarized in *Fig. 1* as the Diamond of Data Professionals.

2. HOW ENTITY/RELATIONSHIP PROFILING ENABLES EACH MEMBER OF A DATA TEAM

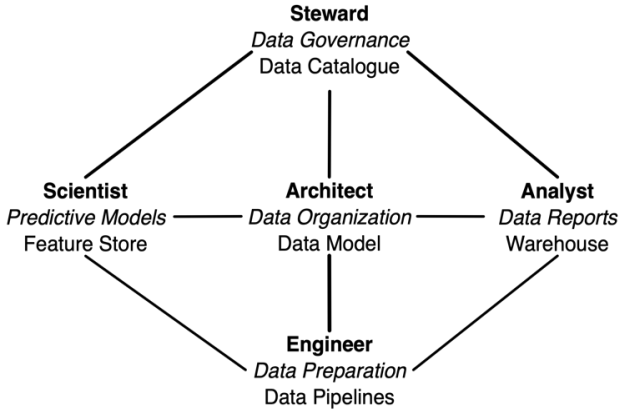

Fig. 1: Shaping the Diamond of Data Professionals with the DataViadotto Profiler

2.1 DATA ENGINEER

A data engineer has the responsibility to prepare data for analysis where either reports or predictive models are developed. Typically, the preparation involves gathering and transforming raw data into structures conformant with the business' models. For that purpose, engineers build and maintain robust data pipelines and monitor how data is being used by applications. Success criteria include the ability of users to access data they need in a timely manner. Engineers implement the models that data architects create, and they also enable the work of data analysts, data scientists, and – in turn – that of data stewards.

Copyright © DataViadotto Limited

Entity/Relationship Profiling is fundamental for the effectiveness of data pipelines that engineers build. Indeed, efficient access to data is implemented by using i) uniqueness constraints and keys that ensure the integrity of business entities and their efficient retrieval by unique indexes, and ii) foreign keys to traverse relationships between business entities across different data sets. Without knowledge of all the keys and foreign keys that hold in a repository, engineers will miss opportunities to understand the relevance of data sets for some application, to retrieve entities or their relationships efficiently, to integrate new data sets within a data pipeline, or to de-duplicate entities or relationships. Since data sets may have large volumes and are often prone to inconsistencies and missing data, the most important business entities and relationships between them may be hidden. It is therefore essential to support Entity/Relationship Profiling from inconsistent and incomplete data at scale, and with an interface where human experts can engage with machine-selected data samples to decide which unique constraints and foreign keys are meaningful for their task at hand. These features of the DataViadotto Profiler are unique across all data profiling tools.

2.2 DATA ARCHITECT

Similar to the data engineer, a data architect also knows how data can be extracted from given sources, how data is transformed into useful formats, and cleaned to meet business requirements. Unlike the data engineer, however, the primary responsibility of a data architect is planning the architecture in which data will be managed. In other words, the architect creates the logical models for data collection, storage, and access for users, while anticipating and adapting to evolving needs of those users. As such, the architect's role is very central: they reflect and inform the entire data governance strategy set out by data stewards, but they directly set out requirements that reports of data analysts and predictive models of data scientists must meet, while the engineer is responsible for implementing and maintaining the data pipeline according to the architect's models.

Entity/Relationship Profiling is central for achieving an actual fit between data assets and the data models that architects create. Since keys and foreign keys constitute the main cornerstones of logical data models, they manifest themselves in classical models such as the relational model of data, but also in feature stores, in data warehouses, and modern architectures such as graph databases. The ability to compute all keys and foreign keys that hold on data assets provides architects with a complete choice of actions that engineers can feasibly take to support the operations that users require. This ability is crucial for the work of stewards, architects and engineers, which can only be effective when they are able to detect drifts in the data, and translate these changes into the underlying models. Hence, continued Entity/Relationship Profiling is a core prerequisite for ``Knowing Your Data'', at any time.

Copyright © DataViadotto Limited

2.3 DATA ANALYST

While data architects and engineers make data available in formats useful for the company, analysts and scientists find facts from data and turn them into actionable insight. A data analyst investigates data from different angles, performs further cleaning and transformations to uncover trends in the data. They may explore new opportunities for their organization to collect additional data for deeper analysis. Data analysts are skilful at mining data, and reporting their results to others, including visualizations and dashboards that communicate their findings at a level the target audience can understand. Typically, data analysts develop performance metrics, report facts about the data, and convey these observations to others, while data scientists need to ensure those observations are actually statistically significant.

Entity/Relationship Profiling is of central importance to the accuracy and transparency of reports that data analysts communicate. Knowledge of all keys and foreign keys allows analysts to choose the best way in which their target audience can relate to the business entities and relationships that appear in their reports. This ensures transparency of their report for the target audience, achieved by business keys and relationships that can be understood and trusted by the audience. Knowledge of all keys and foreign keys helps analysts further clean the underlying data by de-duplicating business entities, such as identifying all the various surrogate ids associated with the same business entity, and curating records that violate referential integrity. In turn, these activities enable data engineers to further boost the performance of running reports. The DataViadotto Profiler employs industry-compliant metrics for ranking the keys and foreign keys that hold on the data. By choosing the metric that matter to them, users are guided towards those keys and foreign keys that may represent the best choice for their task at hand.

2.4 DATA SCIENTIST

Data scientists use statistics and machine learning to move the descriptive analytics of analysts into the realm of predictive analytics by forecasting future events or outcomes. Most of the time, predictive models require a lot of high-quality data to make accurate predictions, and data scientists need to be skilful at selecting the right training data and algorithms in the right context. For example, training data needs to be balanced and represent an unbiased sample of the data. Typically, different predictive models have a trade-off between their accuracy and explainability of their predictions. Data scientists therefore have the responsibility to maximise the portion of possible events that can be predicted with high accuracy using clear explanations of cause and effect.

Copyright © DataViadotto Limited

Similar to other data stores, Entity/Relationship Profiling is also central to the effectiveness and efficiency of feature stores that data scientists manage. An effective feature store understands all the data about business entities relevant for model training or serving a prediction, as well as all the relationships these business entities exhibit across different data assets. For that reason, knowing all ways of identifying business entities uniquely and how they are connected will allow data scientists to manage feature stores effectively and efficiently. Indeed, due to the dependence of predictive models on features (called feature lineage) it is crucial for data scientists to understand how different features are used and what relationships they form across the feature store. Without understanding all the ways entities can be identified uniquely, data scientists cannot ensure that training data is unbiased since the same business entity might have several occurrences in the training data due to different ids being used for it (entity duplication). Likewise, knowledge about all relationships ensures that features are not duplicated within a predictive model, and removing such redundancy ensures the resulting model becomes simpler and more explainable. Similar to reports, Entity/Relationship Profiling facilitates the detection of data and model drifts but within a feature store. And similar to the other data roles, the features of the DataViadotto Profiler ensure that these tasks can be automated as much as possible while ensuring that the human expert in the loop remains in full control and is as informed as possible.

2.5 DATA STEWARD

A data steward keeps track of data assets to ensure they are accessible, functional, safe, and reputable. This covers the entire data lifecycle from creation to usage through storage and deletion. Data stewards are responsible for providing high-quality data that can be readily accessed when needed. They must therefore have the ability to see data beyond silos, identify and maintain business rules that govern the data. Knowing as much as possible about a business' data, keeping them accessible and accurate, data stewards are advocates of gaining a competitive advantage for their company from data by using responsible data analytics, management, science and security.

Entity/Relationship Profiling is fundamental for guiding data stewards in breaking down data silos and maintaining data catalogues that are accurate and comprehensive, all at low cost due to high levels of automation. Indeed, Entity Profiling is the basis for discovering data sets that are relevant for any business entity of interest. Similarly, Relationship Profiling is the basis for discovering how existing data sets are linked, and how new data sets can be integrated into the existing data repository. Using profiling to provide as much automation as possible will reduce the effort and resources required to keep data catalogues accurate and up to date.

Copyright © DataViadotto Limited

3. CLOSING

In summary, *Entity/Relationship Profiling* is of fundamental importance to the effectiveness and efficiency of any data role. The distinctive features of the *DataViadotto Profiler* enable data stewards, architects, engineers, scientists and analysts to fulfil the vision of their role and contribute towards data stories about business entities and their relationships that will give your organization a competitive advantage, and lead to satisfied employees and customers. Hence, the *DataViadotto Profiler* is the only choice to shape your diamond of data professionals within your organization.

ABOUT DATAVIADOTTO

DataViadotto is the industry pioneer for Entity/Relationship profiling technology. The company draws on decades of academic research in the subject to make the process of discovering models from data more effective, efficient and intuitive. Ultimately, data becomes profitable.

FOR ADDITIONAL QUESTIONS, CONTACT DATAVIADOTTO

www.viadotto.tech

Copyright © DataViadotto Limited