WHITE PAPER

The Return on Connected Data

More Connections, More Value

Dave Packer, Product Team

White Paper

TABLE OF CONTENTS

Connections Unlock the Value of Data 1

Data, Data Everywhere, Nor any Drop to Drink 1

The Power of Connected Data 2

Quench Your Thirst for Insights with Connected Data 3

Neo4j: Your Path to Connected Data 4

Conclusion 4

WHAT IS CONNECTED DATA?

Connected data is the representation, usage and persistence of relationships between data elements. The key here is to maintain ongoing knowledge of the relationship and not simply instantiate it as might be done using a JOIN table in a relational database management system (RDBMS).

The Return on Connected Data

More Connections, More Value

Dave Packer, Product Team

Connections Unlock the Value of Data

The value of data comes from an organization's ability to understand it in relation to other data. By itself, data offers finite value. When connected, data's value is infinite.

Case in point: There's a difference between knowing that a customer bought a child's winter coat and knowing that a customer bought a child's winter coat for the last three years, all in the color blue, in successive sizes.

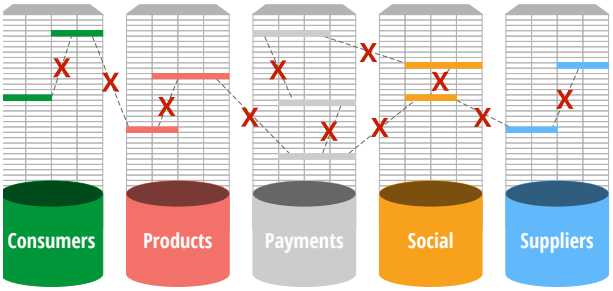
Increasing data's connectedness further increases its value through additional context. This is another demonstration of Metcalf's Law of the Network: network value increases exponentially as you add users or nodes to the network.

Enticed by the promise of valuable business insights, companies have invested in big data technologies. But dumping data in a data lake doesn't preserve nor reveal the relationships between data points. And JOIN tables only materialize a relationship when the query is run; relationship information is not a first-class entity in relational nor other types of databases.

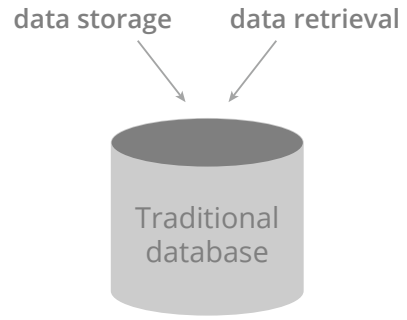
Yet organizations express the need for connected data everywhere, especially when they are connecting people — like employees or customers — to products, to business processes, to networks and computers and Internet-enabled things.

Because relational databases don't persist relationship information in storage or any other stage of their analytic exercises, finding connections requires an enormous amount of extra processing. And persisting these connections over their lifetime is next to impossible.

This white paper shows business leaders how to take advantage of relationships that either live in data or are represented by data as part of how people interact with one another or with processes, and how these processes interact with systems and devices.

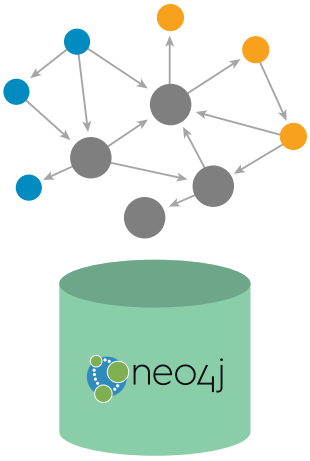


Data stored in relational databases (RDBMS) or other data silos fails to leverage how the data is connected across the organization.



Finding connections takes another tool and considerable processing work. Even then, connections are not stored

VS.



Neo4j stores connections with the data. There's no extra work: it's already connected

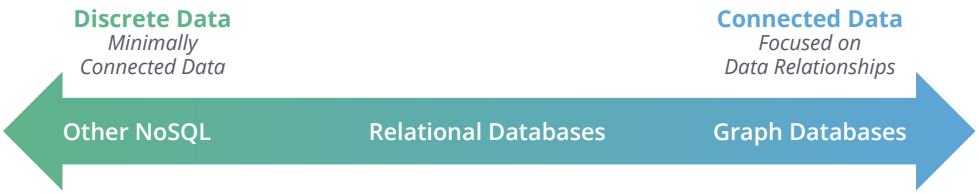
Data, Data Everywhere, Nor any Drop to Drink

Like the ancient mariner at sea, organizations are surrounded by data but they're limited in their ability to do anything with it.

They've collected all the data they could get their hands on in a desire to uncover the business insights that big data is known for. But despite having plenty of data, and data integration tools, organizations are failing to realize big data's value. That's because the value isn't in disparate data but in the relationships, or connections, between the data — and this information is not easily available.

Data doesn't become connected by simply dumping it into a central data lake. Data becomes connected when you treat relationship information as a first-class entity — persisting it, assigning properties to it, and using it as a means to develop context for applications. Thus, it's best to think of big data needing to be ***connected***, not merely ***contained***.

Without connected data, organizations lack key information that's necessary to obtain a 360-degree view of a customer, build a complete network topology, deliver relevant recommendations in real-time, or obtain the visibility required to prevent fraud. The data landscape has grown, as has the number and types of connections upon which this data depends, making the need to surface connected data much more urgent.



As connected data becomes more essential to the success of an application or project, the role of graph database technology becomes more of an imperative.

CONNECTED DATA ENABLES ORGANIZATIONS TO:

- Gain a 360-degree view of a customer
- Build a complete network topology
- Deliver relevant recommendations in real-time
- Obtain the visibility required to prevent fraud
- Introduce data science and AI workflows to applications

The Neo4j graph contains both the product catalog and the attributes of shopper interactions, creating and storing connections between both datasets as shoppers have more conversations.

CASE STUDY:

eBay, Inc. is continually looking to improve the way shoppers find the perfect item since the typical search box experience regularly falls short in understanding and remembering what a shopper is truly seeking to find.

As an example, SVP & Chief Product Officer RJ Pittman considers the query: "My wife and I are going camping in Lake Tahoe next week, we need a tent."

Most search engines would focus on the word "tent." But the additional context regarding location, temperature, tent size, scenery, etc. is typically lost. Yet, this type of specific information is actually what informs many buying decisions. Relaying or maintaining this context (i.e., connected data) is often a burden left to the user and a new solution was needed to remove the hard work associated with shopping.

To remedy this issue, eBay used a Neo4j knowledge graph to create the eBay ShopBot: a smart, personal shopping "bot" that converses with users via text, voice or photo search capabilities while parsing these conversations for meaning and context. This required a tool that could both efficiently navigate and leverage vast amounts of connected data.

To build the eBay ShopBot, the engineering team needed a robust knowledge graph of connected data in addition to natural language understanding and artificial intelligence to store, remember and learn from past interactions with shoppers.

eBay chose Neo4j as the native graph database that holds the probabilistic models that aid understanding in the conversational shopping scenario. The Neo4j graph contains both the product catalog and the attributes of shopper interactions, creating and storing connections between both datasets as shoppers have more conversations.

When a shopper searches for "brown bags" for example, eBay ShopBot knows what details to ask about next, such as type, style, brand, budget or size. As it accumulates this information by traversing through the graph, the application is continuously checking inventory for the best match.

The eBay team expects to deploy the chatbot across multiple platforms via plugins including Slack, Microsoft and Facebook Messenger. The initial plugin operates in Facebook Messenger and is available at https://shopbot.eBay.com/.

The Power of Connected Data

The biggest benefit of connected data is the ability to provide a connected view of the data to your analytic and operational applications, thereby gaining and growing intelligence downstream. The connections can be made available or revealed to applications or business users to make operational decisions.

You also obtain context that allows you to more deeply or better refine the pieces of information you're collecting or the recommendations you're producing. Marketing may determine the best time to send an email to customers who previously purchased winter coats and dynamically display photos in their preferred colors. The more understanding you have of the connectedness or of the relationships between data, the better and more refined your system is downstream.

Connected data applies to a variety of contexts. In addition to refining the output of your recommendation engines, you can better understand the flow of money to detect fraud and money laundering, and assess the risk of a network outage across computer networks. Connected data also helps you see when and how relationships change over time. For example, you can determine when a customer moves and change the applicable data (such as mailing address) so that customer data doesn't become obsolete and useless to applications.

Connected data is most powerful when it provides operational, real-time insights and not just after-the-fact analytics. Real-time insights allow business users and applications to make business decisions and act in real-time. Thus, recommendation engines leverage data from the current user session to deliver highly relevant options that the customer wants or needs at that moment.

IT organizations can proactively mitigate network issues that would otherwise cause an outage, and the antifraud team can put an end to potentially malicious activity before it results in a substantial loss.

Discovering connections between data points is essential to detecting first-party bank fraud at financial institutions. If customer, loan, credit card and contact data are all reviewed discretely, finding such sophisticated fraud patterns wouldn't be possible.

Connected data is most powerful when it provides operational, real-time insights and not just after-the-fact analytics.

Connected data thus provides the necessary linkages between silos of data components and provides an understanding of the overall data landscape.

CASE STUDY:

With over 3500 employees located across 20 global offices, Airbnb is growing exponentially. As a result of employee growth, they have experienced an explosion in both the volume and variety of internal data resources, such as tables, dashboards, reports, superset charts, Tableau workbooks, knowledge posts, etc.

As Airbnb grows, so do the problems around the volume, complexity, and obscurity of data. Information and people become siloed, which creates inefficiencies around navigating personalized tribal knowledge instead of clear and easy access to relevant data.

In order for this ocean of data resources to be of any use at all, the Airbnb team would need to help employees navigate the varying quality, complexity, relevance and trustworthiness of the data. In fact, lack of trust in data was a constant theme because employees were afraid of accidentally using outdated or incorrect information. Rather, employees would create their own additional data resources, further adding to the problem of myopic, isolated datasets.

To address these challenges, the Airbnb team created the Dataportal, a self-service system providing transparency to their complex and often-obscure data landscape. This search-and-discovery tool democratizes data and empowers Airbnb employees to easily find or discover data and feel confident about its trustworthiness and relevance.

When creating the Dataportal, the Airbnb team realized their ecosystem was best represented as a graph of connected data. Nodes were the various data resources: tables, dashboards, reports, users, teams, etc. Relationships were the already-present connections in how people used the data: consumption, production, association, team affinity, etc.

Using a graph data model, the relationships became just as pertinent as the nodes. Knowing who produced or consumed a data resource can be just as valuable as the resource itself. Connected data thus provides the necessary linkages between silos of data components and provides an understanding of the overall data landscape.

Given their connected data model, it was both logical and performant to use a graph database to store the data. Using Hive as their master data store, Airbnb exports the data using Python to create a weighted PageRank of the graph data before pushing it into Neo4j where it's synced with Elasticsearch for simple search and discovery.

Quench Your Thirst for Insights with Connected Data

That all begs the question: How does data become connected? Don't worry, you don't have to boil the ocean. Connecting data is easy and approachable. In fact, you don't even have to throw out what you're already doing.

A graph database makes it easy to express and persist relationships across many types of data elements (people, processes, and things). A graph database is a highly scalable transactional and analytic database that stores data relationships as first-class entities.

Think of a graph data model as a mind map composed of two elements: A node and a relationship. Each node represents an entity, and each relationship represents how two nodes are associated. By assembling the simple abstractions of nodes and relationships into connected structures (imagine connecting two circles with an arrow or line), a graph database enables you to build sophisticated models that map closely to a problem domain.

Graph databases are simple and agile due to their schema-optional nature. Because the structural overhead of traditional database schema is eliminated, graphs can be easily changed or updated. You aren't required to have a similar structure for every node. You can put data into a graph simply by reading in a table, and if you want to add another column or attribute to that data element, you simply add the attribute.

Graph technology and connected data are being used today. In fact, companies like LinkedIn and Google recognized early on the need to use the relationships in their data. Fortunately, graph databases have become more accessible. Mainstream graph technology makes it easy for any company to apply the concept of connected data to more problems.

Regardless of where you are today in your big data initiative, a graph database can help you reach the next level of maturity. If you're just starting out, a graph database can address the challenge of migrating all of your data into one connected place. There are no data typing requirements or fixed structure for how the graph is supposed to be defined. The graph can automatically determine the data typing, and additional nodes can be added at any time.

You can also get more value out of your big data technology by using graphs to transfer knowledge of what the organization has done across other departments. Graphs keep relationship information always at hand, but that information changes over time. The graph can therefore become a longstanding system of record that delivers value to different departments. For example, it can enable the customer service department to approve or deny a product exchange based on the customer's purchase, return, and exchange history.

By connecting data, graphs also enable you to leverage all of the data you've dumped into your data lake. You can simply port one table or CSV file at a time into the graph database, and the database will make nodes out of the rows in the tables.

By looking at how the tables join together and building relationships from that information, you can create a graph that links discrete data, such as data from SAP to Oracle to Salesforce to Marketo to the bespoke shopping cart application, and from there you can begin to reveal the connections that a data lake doesn't expose.

Identifying connections in your data is only helpful if you know what to do with them. Graphs can help you with that, too. Today's applications, whether they be recommendation engines or anti-money laundering systems, are powered by graphs. The graph database operationalizes the data by making it possible for the app to traverse the data. Once you have this ability, the data becomes actionable.

Regardless of where you are today in your connected data initiative, a graph database can help you reach the next level of maturity.

By tracking patterns of connected data around when users leave and arrive home, the Telia Zone offers highly accurate and relevant suggestions regarding when users will most likely be at home – information that is both elusive and salient to every smart home service.

CASE STUDY:

Telia is the incumbent telecommunications carrier in the Nordic market across mobile, broadband, consumer and enterprise markets. In order to stay competitive, the Telia team decided to reinvigorate an area of least innovation: their consumer broadband business.

Taking a closer look, the Telia team recognized their roughly 1 million installed Wi-Fi routers as underutilized assets. Their goal was twofold: help consumers simplify their lives and help consumers better entertain within their homes.

The solution: The Telia Zone. After tracking all Wi-Fi connected devices within their customers' homes and recording when a device leaves or enters (all collected with prior consent), the Telia team provides this connected dataset and a number of APIs to a partner-driven ecosystem of apps and services. The result is a smart home platform that allows consumers to pick and choose apps that meet the two goals of simplification and home entertainment.

Telia Zone apps do everything from reminding consumers when they forget to lock the door after they leave for the day to letting them know when their kids arrive home after school (and turning on the lights when they get there). Another example is the Coplay app which automatically generates a Spotify playlist based on which party guests are connected to a home's Wi-Fi. If a guest leaves, their songs are dynamically removed from the playlist; and when they return, so do their favorite songs.

By tracking patterns of connected data around when users leave and arrive home, the Telia Zone also offers highly accurate and relevant suggestions to other apps regarding when users will most likely be at home – information that is both elusive and salient to every home delivery service.

Powering it all on the backend is the Neo4j Graph Platform. The Telia team chose Neo4j because their smart home dataset was already a graph of connected data. Also, with Neo4j's Causal Clustering architecture, the Telia team was able to horizontally scale their operations without always knowing what given traffic loads might be. Furthermore, the graph data model allows them the flexibility to keep evolving the Telia Zone platform without breaking existing components.

The Return on Connected Data

WHY NEO4J?

Native graph store
Unlike other database technologies, Neo4j is designed from the ground up to store and retrieve data and its connections. Relationships are first-class entities in a native graph database – making them easier to query and analyze.

Complete graph platform
Neo4j offers a complete graph platform that includes a native graph database, analytics, ETL and graph visualization.

Flexible schema
Neo4j's versatile property graph model makes it easier for organizations to evolve solutions as data types and sources change.

Performance and scalability
Neo4j's native graph processing engine supports high-performance graph queries on large user datasets to enable real-time decision making.

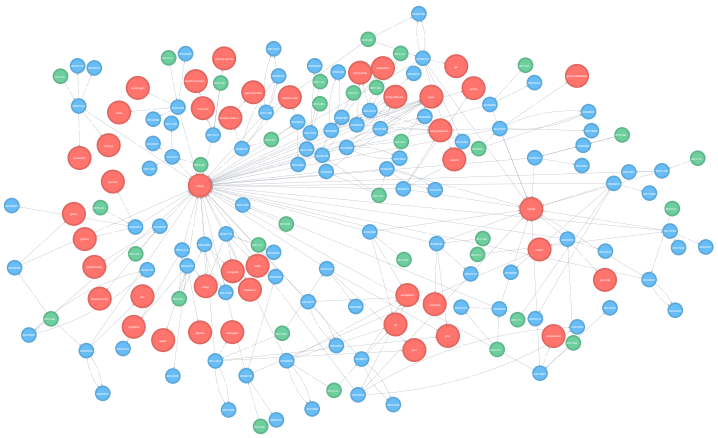
High availability
The built-in, high-availability features of Neo4j ensure your user data is always available to your mission-critical next generation service assurance solution.

Neo4j: Your Path to Connected Data

Neo4j offers an innovative, reliable native graph platform that reveals and maintains the integrity of connected data from the moment it's conceived through each stage of design, development, analysis, implementation, and operation. Neo4j offers this connections-first approach as a new paradigm to reveal and utilize relationships among data in order to help organizations along their path to new generations of applications and analytic advancements.

Not only does Neo4j make connected data possible; it makes connected data ***easy***. As graphs are built out of your data, Neo4j automatically creates connections as they exist and reveals connections that you may have never considered.

For example, you may discover that a prospect's address is within driving distance of a sales rep. That connection requires multiple hops, but the graph database allows you to do so, going deeper into your data and identifying the shortest path to get there. Neo4j also allows you to move beyond connecting data for analytic purposes to connecting business processes and the people involved in those processes as well.



Neo4j allows you to easily query and visualize your connected data for patterns and trends that are essential for a sustainable competitive advantage.

Conclusion

With all the data available today, no organization should be thirsty for business insights. But in order to obtain the deepest and most revealing insights, you must understand the connections between your data and preserve relationships as you think of them, draw them, query them, or store them on disk.

The difference in value is like moving from two dimensions into five, or from black and white to 3D technicolor. With Neo4j, enabling connected data is easy. Organizations can use all their data more effectively, improving data quality, driving revenue through real-time recommendations, preventing fraud, and running their operations smoothly and securely.

Neo4j quenches the thirst for connected data, connecting people to products to processes. If you'd like to learn more about Neo4j, contact a Neo4j expert or try out the free Neo4j Sandbox.

Neo4j is the leader in graph database technology. As the world's most widely deployed graph database, we help global brands – including Comcast, NASA, UBS, and Volvo Cars – to reveal and predict how people, processes and systems are interrelated.

Using this relationships-first approach, applications built with Neo4j tackle connected data challenges such as analytics and artificial intelligence, fraud detection, real-time recommendations, and knowledge graphs. Find out more at neo4j.com.

Questions about Neo4j?

Contact us around the globe:
info@neo4j.com
neo4j.com/contact-us