BREAK FREE FROM THE SHACKLES OF ETL AND ELT

HOW A NEW DATA ENGINEERING APPROACH IS TRANSFORMING THE MODERN ENTERPRISE ANALYTICS LANDSCAPE

incorta

WHAT'S HAPPENING

Inefficiencies surrounding data transformation have long plagued IT teams, who spend countless hours navigating (or simply hacking through) the thorny limitations of traditional data infrastructure. The majority of IT teams today still rely on processes like ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform)—conceived in the 1990s and late 2000s, respectively—to map the connections between data sources and make the data available for analytics.

Like frogs in boiling water, many IT teams today have become so accustomed to the harsh realities of data transformation that they no longer realize the impact it is having on their organization. After decades without relief, the costs associated with these outdated processes are now widely accepted as the cost of doing business.

But it doesn't need to be this way. We are now seeing IT teams embrace a new data engineering approach that provides immediate access to their organization's data without the drawbacks of ETL or ELT. By making any database (or data source) continuously available for analytics, this new approach completely changes the way a company interacts with its data.

That's an absolute must in today's fast-paced business environment, as soaring demand for data and self-service enterprise analytics pushes the limits of what IT can deliver. Business users want more access to more data, and they no longer have the luxury of waiting days or weeks for IT to provide dashboards and reports.

This trend shows no signs of slowing anytime soon. The accelerating rate, volume, and complexity of enterprise data being generated today demands that organizations find a better, simpler, and more efficient way. (share this 🐦). Modern business moves fast—and it's time for data and analytics technology to catch up.

To illustrate how this new approach to data engineering transcends the limitations of its ETL and ELT predecessors, let's examine how data flows through each of these transformation processes, unpacking the various complexities IT must navigate along the way.

Direct Data Mapping gives IT teams a new model for data processing that eliminates the need for ETL and ELT, and gives organizations immediate access to their data. 🐦

THE PROS AND CONS OF 3NF DATABASES

The majority of analytics solutions draw from a similar type of source: a database that stores data in the Third Normal Form (3NF). This model hasn't changed much in decades—and that's because it works pretty well. 3NF databases are built to support fast-moving organizations. They make data entry easier, allow for multiple users to update the database concurrently, and improve an organization's overall data quality by eliminating redundant or duplicate data.

The 3NF database is not without drawbacks, however. As the number of tables and joins in the database grows, things can get out of hand quickly. Because they rely on hundreds—if not thousands—of separate tables to keep data organized, the relational topography can easily lead to runaway complexity and paralysis. What's more, analytical and operational queries don't work especially well in highly-complex environments, and when they do, the system moves very slowly. Any multidimensional question necessitates jumping around between dozens or hundreds of disparate tables just to gather the relevant data and see how it all fits together.

Because present-day 3NF databases do not lend themselves to analysis, additional steps are needed to "bridge the gap"—i.e. make the data in 3NF friendlier and more usable for complicated, multidimensional queries. **This is, of course, the exact reason why ETL and ELT processes exist.**

Third Normal Form (3NF) Data Model

A common starting point for the data transformation process. While ETL, ELT, and DDM typically start with a 3NF data model like this one, the similarities end here.
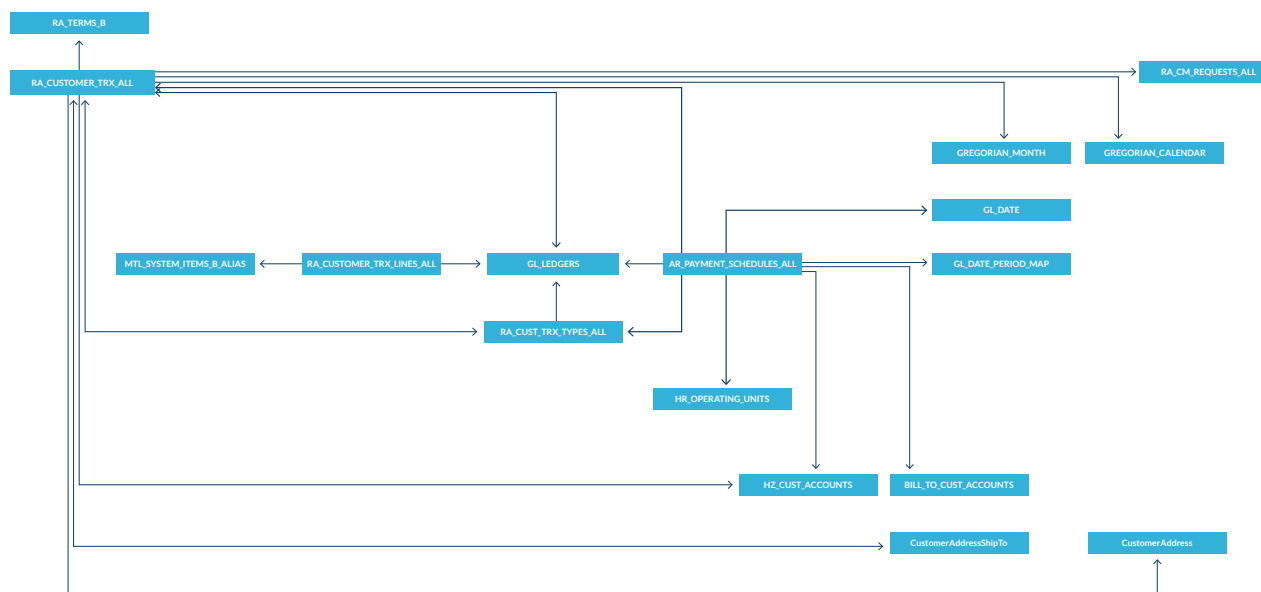

Fig. 1

EXTRACT, TRANSFORM, LOAD (ETL): MODERATE FIDELITY, LOW SPEED

With ETL, the first step is to **extract** all of the data into an environment where it can be worked on—typically an ETL tool. For a database with hundreds or thousands of tables, each with tens of thousands of entries, this can take a long time—sometimes dozens of hours, or even days.

After extracting the data at a point in time, the data can then be **transformed** into a more simplified format against which analytics can be run. In this phase, which happens inside the ETL tool, the data is effectively "flattened" and standardized in preparation for being passed onto analytics platforms.

Reorganizing and assembling the data in an analytics-ready format requires significant computing power and adds even more time to the ETL process. The transformation process also involves data enrichment, with the language and values surrounding the data being updated to more closely reflect terms that can be understood by the business user. This could be as simple as translating a series of letter codes on sales orders (e.g. "W/S/A") into business-relevant statuses (e.g. awaiting delivery, shipping, at warehouse, etc.), allowing business users to better understand the data and analytics that are returned.

Finally, the data needs to be **loaded** into its final destination, often (but not necessarily) a data warehouse. From there, the data feeds into the organization's analytics tool of choice. In this stage, we encounter our old nemesis, the star schema, whose organizational structure helps to ensure the data is in the optimal form for digestion by the analytics layer and that the correct subset of data is available to answer other specific questions a team might have. This is often the most time-consuming leg of the ETL process.

Extract, Transform, Load (ETL)
Introduced in the 1990s. Built for a world where business questions rarely change.

Fig. 2

What does all of this mean for business teams? If the organization is starting from scratch, it can take months or even years to build an ETL model and get answers to their questions. In that time, the team's questions may have changed and new data columns may be needed to answer them. Because insights are so rigidly tied to the ETL graph, making these updates requires changes to the underlying model, which means weeks of waiting— essentially, a never-ending cycle of ETL.
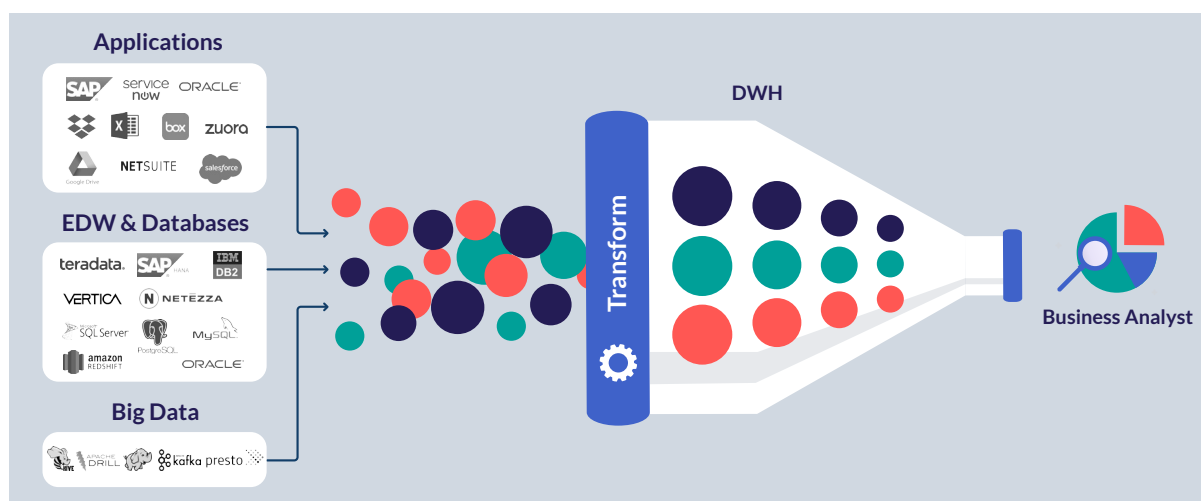
Even when ETL graphs do succeed, considerable maintenance is required to keep them up to date with the latest data. These updates can take up to half a day to process just a single day's worth of transactions.

Running ETL is a complex operation that involves multiple transformations of the data. It prevents you from becoming agile and responsive because you are left operating your business on snapshots of the world taken every six hours.

> **"There were a number of applications out there that were processing data fast, but all of them relied on data simplified by ETL."**
>
> **ALAN DAVIDSON,**
> DIRECTOR IT INFRASTRUCTURE, BROADCOM

A Classic ETL Graph

A partial real-world ETL graph from a Fortune 1000 company. Each of the boxes below represents a transformation. Tens, if not hundreds, of similar, tightly-linked graphs are required to complete the entire ETL process.
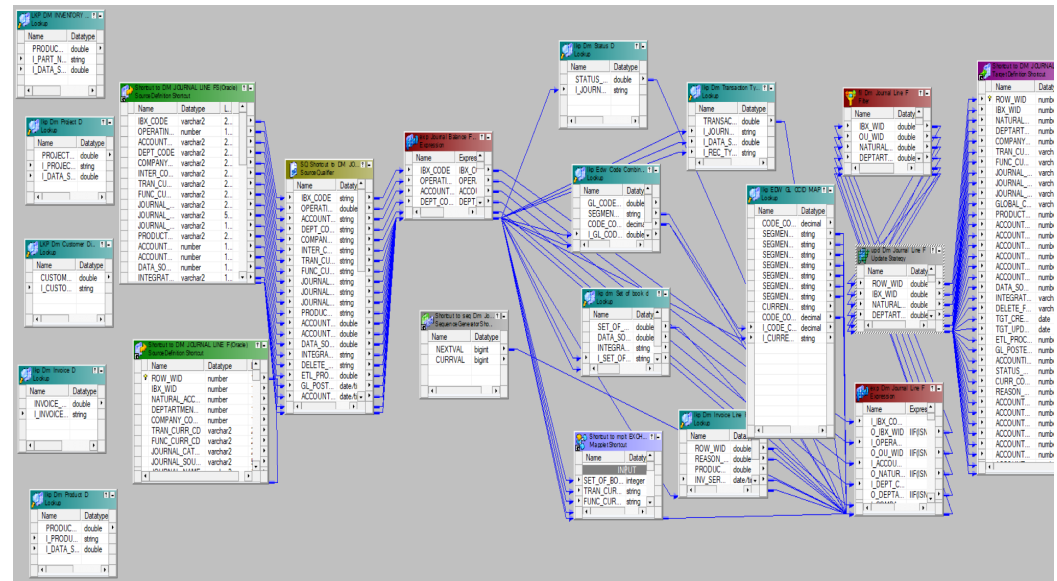

Fig. 3

As you can see, ETL graphs are brittle and full of interdependencies. Making even the slightest change to any part of the ETL graph results in cascading and unpredictable effects throughout the rest of the system.

EXTRACT, LOAD, TRANSFORM (ELT): FASTER RESULTS, BIGGER DEMANDS

In an effort to overcome the limitations of ETL, many IT teams have switched to a slightly newer approach to data processing that takes advantage of the availability of fast and cheap computing power: Extract, Load, Transform (ELT). Made possible in great part by the rise of cloud-based "analytical databases," ELT makes it easier to process larger amounts of data.

Extract, Load, Transform (ELT)

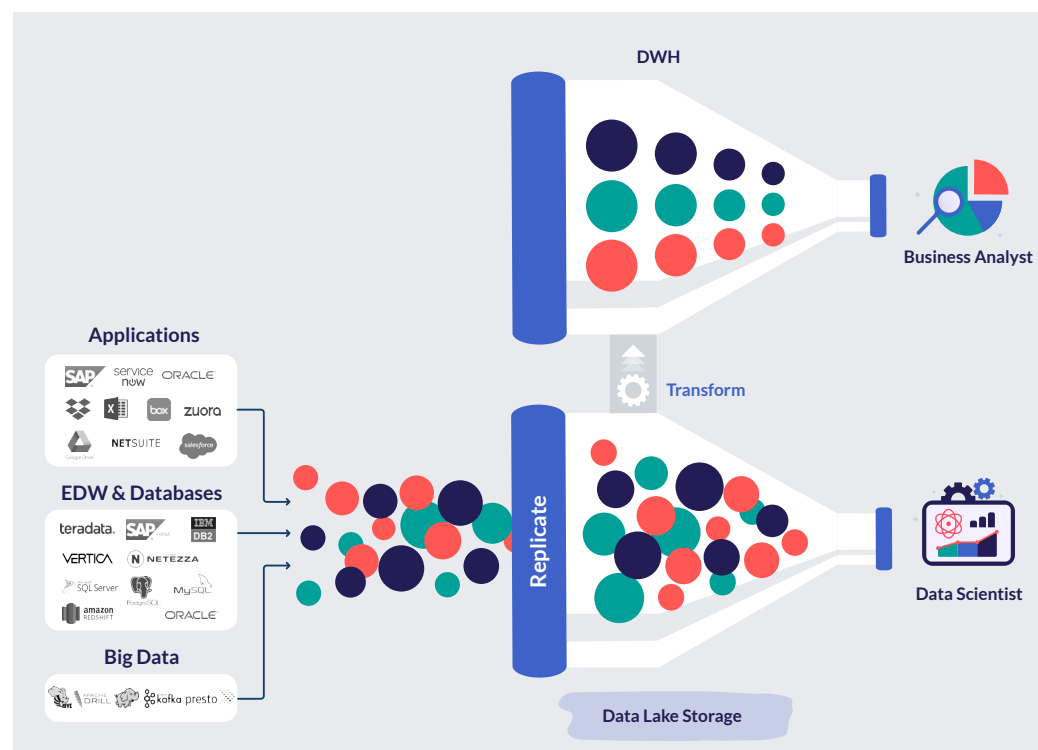Introduced in the late 2000s. Built for speed and flexibility, but hard limits remain.

Fig. 4

As we saw in the chart above, the ETL process can become complex very quickly and can even necessitate "ETLs within ETLs" as complexity grows. ELT sidesteps this problem by essentially boiling down these multi-step processes into singular (albeit still complex) database commands.

In the **extract** phase, the data moves out of the source database—mostly unchanged—and loads into a database or data lake, where it is then transformed into a star schema for analysis. Instead of running transformations before loading the data into a data warehouse, you load the data first—and you **load** it into a data warehouse or data lake that is optimized for data transformations (hence "E-L-T").

Finally, the data is **transformed** and the final preparations for analytics happen right inside the data warehouse. This typically gives IT teams more speed and agility because the data no longer has to go through a transformation process before being loaded into a data warehouse. Does ELT mean that analytics dashboards will deliver insights faster? Not exactly. Analytics still cannot be run directly against the data—star schemas are once again required to translate it into a more analytics-ready format and the database must be powerful enough to handle those transformations. In the past, this meant spending big bucks on computing power—meaning you got what you paid for. Today, as cloud computing becomes ubiquitous, ELT and its out-sized appetite for raw computational horsepower are indeed becoming much more viable for many organizations. ELT has proven to be a more flexible approach than ETL.

But, because ELT processes still require star schemas or some other analytics-ready data shape (e.g. cubes and marts), any business team will still need to have a rather specific idea of the questions they want to ask, and the IT team will have to find the types of data needed to answer them. Queries still run very slowly, and the size of databases can quickly balloon out of control. This could mean waiting weeks for insights.

It is also important to note that ETL and ELT are not mutually exclusive approaches to processing the deluge of data that today's organizations are processing. In fact, many find that a combination of both works for them, and so there are many flavors of hybrid ETL/ELT to consider—but at the end of the day, it still means attempting to get flexibility out of an inherently rigid model.

"We set aside four hours for the initial proof-of-concept meeting with Incorta. But in just one hour, we realized that not only would we be saving thousands of person-hours on ETL processes but also had access to a whole new view of our data that we didn't know was possible. In that hour, we were able to analyze tens of millions of transactions. That's something that usually takes weeks."

IT DIRECTOR,
TOP 10 UNIVERSITY

DIRECT DATA MAPPING (DDM): FULL AVAILABILITY, FRESH PERSPECTIVES

Remember that unwieldy 3NF database, the source of all this data that eventually turns into business intelligence? ETL and ELT processes ultimately force you to shape the data into a rigid model against which analytics can be run.

A unified data analytics platform powered by Direct Data Mapping changes all of that. DDM can deal with the data in its original form and makes the 3NF data model (or any kind of database) immediately available for analytics. 🐦 While DDM of course allows for data enrichment, it does not depend on the types of transformation processes required to make traditional methods like ETL and ELT work. It takes 3NF from being the first phase in a long process of becoming analytics-ready to being immediately analytics-ready.

Direct Data Mapping (DDM)

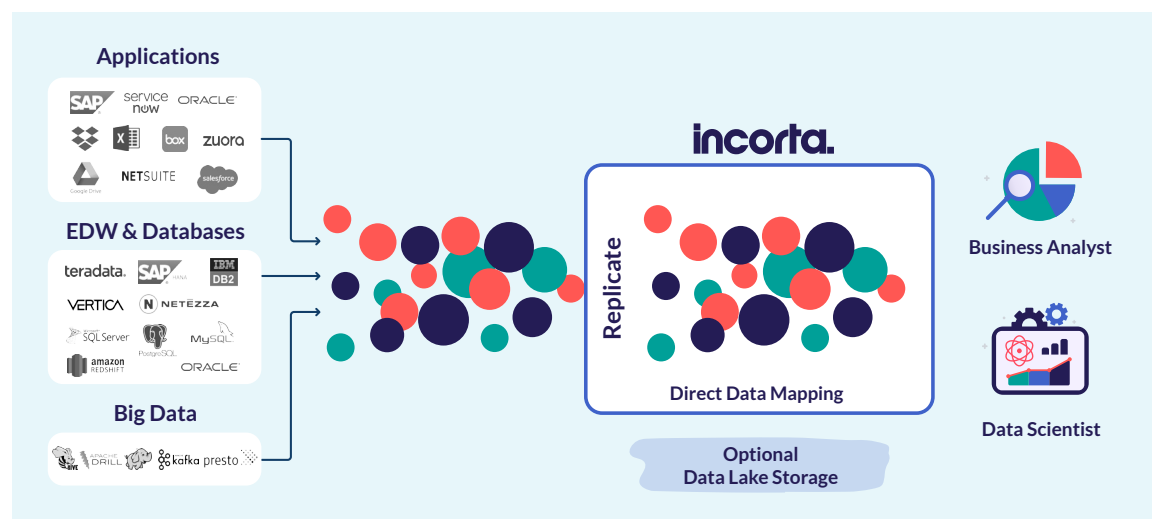Introduced in the late 2010s. Delivers unmatched speed, flexibility, and fidelity.

Fig. 5

DDM leads to orders-of-magnitude improvements in the speed, fidelity, and accessibility of insight. 🐦 Business users can pull up a dashboard and find all of the data they possibly need to answer questions.

They can ask follow-up questions and get answers in subseconds, connect the dots on data from disparate sources, and share findings with others throughout the organization. No longer constrained by lag time, everyone can make more effective "real-time" decisions.

With DDM, enterprise analytics stops reflecting a weeks-old reality and accelerates to the "speed of thought." Business teams can go directly to their analytics dashboard and pull up any and all data they need, no longer needing to predefine the scope of their query. With DDM, they gain the flexibility to consult any data source and easily combine it with what is already available.

The end result can be transformative for business. Queries that used to take hours or days can now happen in subseconds. At long last, enterprise analytics can evolve from a guided tour to a true exploration of the factors that make the business run, or slow it down. When a

question takes a week to answer, it's no surprise that even the most curious employees stop asking questions. The ability to continually ask questions in near real time allows the business to be truly inquisitive and change their lines of questioning as quickly as business conditions change.

With DDM, enterprise analytics stops reflecting a weeks-old reality and accelerates to the "speed of thought."

DDM REDEFINES YOUR RELATIONSHIP WITH DATA

Across every industry today, businesses are under constant pressure to reinvent themselves. Anyone working with technology is looking for a way to make a leap and not just a step; a way to turn linear progress into exponential progress.

Extract, Transform, Load worked well in environments where the questions weren't changing. It was the first way data scientists found to get answers to complicated questions. It was expensive and error prone—but it returned real answers. Companies today, however, need to answer questions quickly, and those questions may change on a daily or even hourly basis. Now that business and the world in general are moving so much faster, its days are numbered.

Extract, Load, Transform was the next step forward. It offers an upside in speed and agility while demanding a larger share of raw computing horsepower. It may reduce the time it takes to get data to the dashboard—but won't necessarily make the dashboard provide answers any quicker. For companies with the scale and capital to make investing in powerful analytical databases worthwhile, ELT can be sufficient for the company's most basic data needs. Yet even with the computing power to crunch it all, business teams will still be constrained by the limitations of the star schema. Achieving fast, agile, and reliable self-service enterprise analytics will remain a near-insurmountable challenge at the company level with ELT.
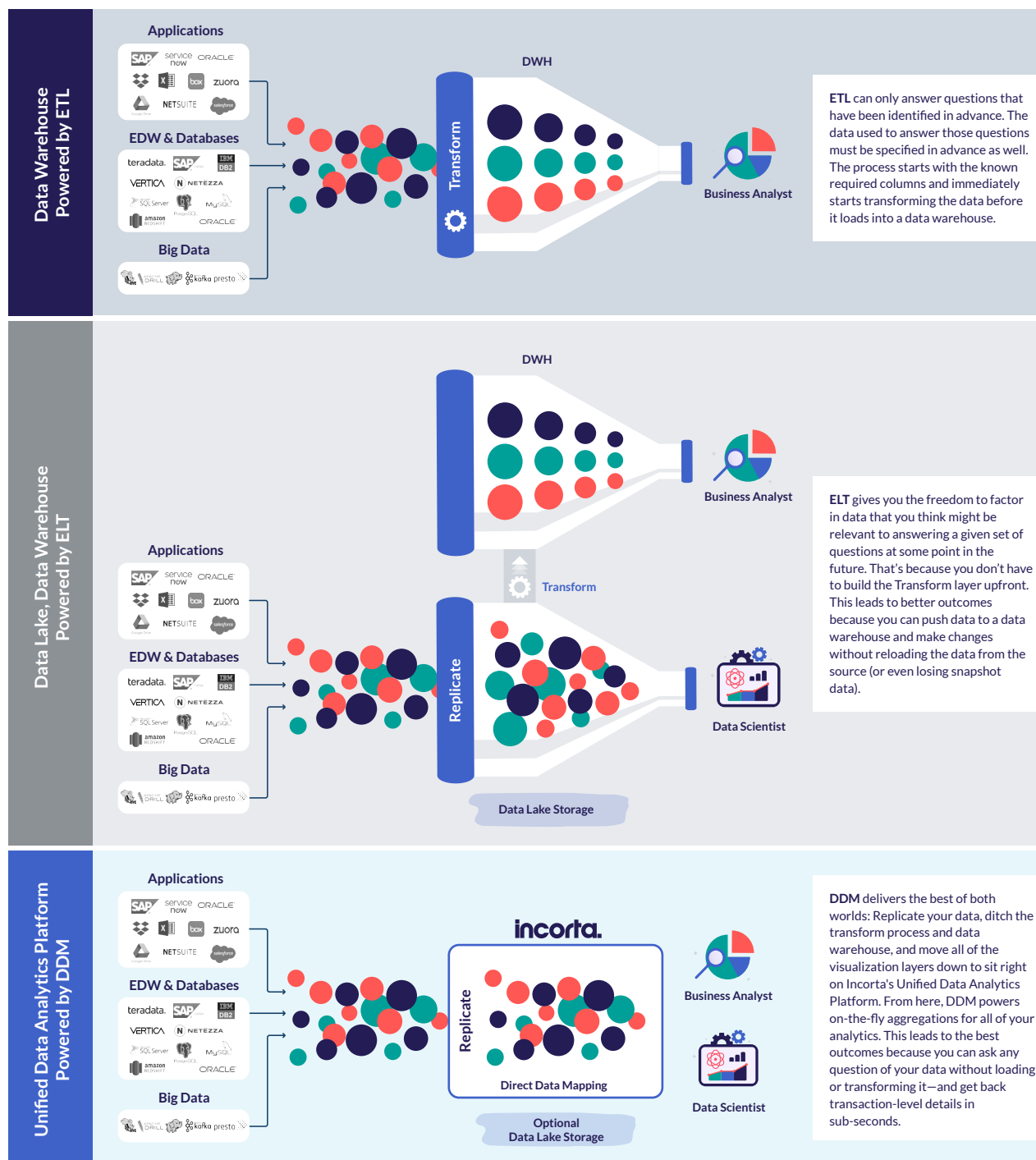
Incorta's **Direct Data Mapping** is a quantum leap in data analytics. It makes data available for analytics in a fraction of the time, it improves the speed of those analytics, and it allows business users to get to answers on a whole new time scale measured in subseconds—not

weeks. In today's highly-competitive business environment, it is no longer enough to make gradual, incremental advances in time-to-insight—the exponentially increasing speed of modern business means organizations must be looking for entirely new ways to make their unruly data depots as available as possible for analytics and data-driven decision making.

Questions or comments? Send them my way on Twitter using the hashtag #ETLShackles or ping me directly at @layereddelay

BRINGING IT ALL TOGETHER: A SIDE-BY-SIDE COMPARISON

Data Warehouse Powered by ETL

Applications

EDW & Databases

Big Data

DWH

Transform

Business Analyst

ETL can only answer questions that have been identified in advance. The data used to answer those questions must be specified in advance as well. The process starts with the known required columns and immediately starts transforming the data before it loads into a data warehouse.

Data Lake, Data Warehouse Powered by ELT

DWH

Business Analyst

Applications

EDW & Databases

Big Data

Replicate

Transform

Data Scientist

Data Lake Storage

ELT gives you the freedom to factor in data that you think might be relevant to answering a given set of questions at some point in the future. That's because you don't have to build the Transform layer upfront. This leads to better outcomes because you can push data to a data warehouse and make changes without reloading the data from the source (or even losing snapshot data).

Unified Data Analytics Platform Powered by DDM

Applications

EDW & Databases

Big Data

incorta.

Replicate

Direct Data Mapping

Optional Data Lake Storage

Business Analyst

Data Scientist

DDM delivers the best of both worlds: Replicate your data, ditch the transform process and data warehouse, and move all of the visualization layers down to sit right on Incorta's Unified Data Analytics Platform. From here, DDM powers on-the-fly aggregations for all of your analytics. This leads to the best outcomes because you can ask any question of your data without loading or transforming it—and get back transaction-level details in sub-seconds.

THE DIRECT DATA PLATFORM™

incorta