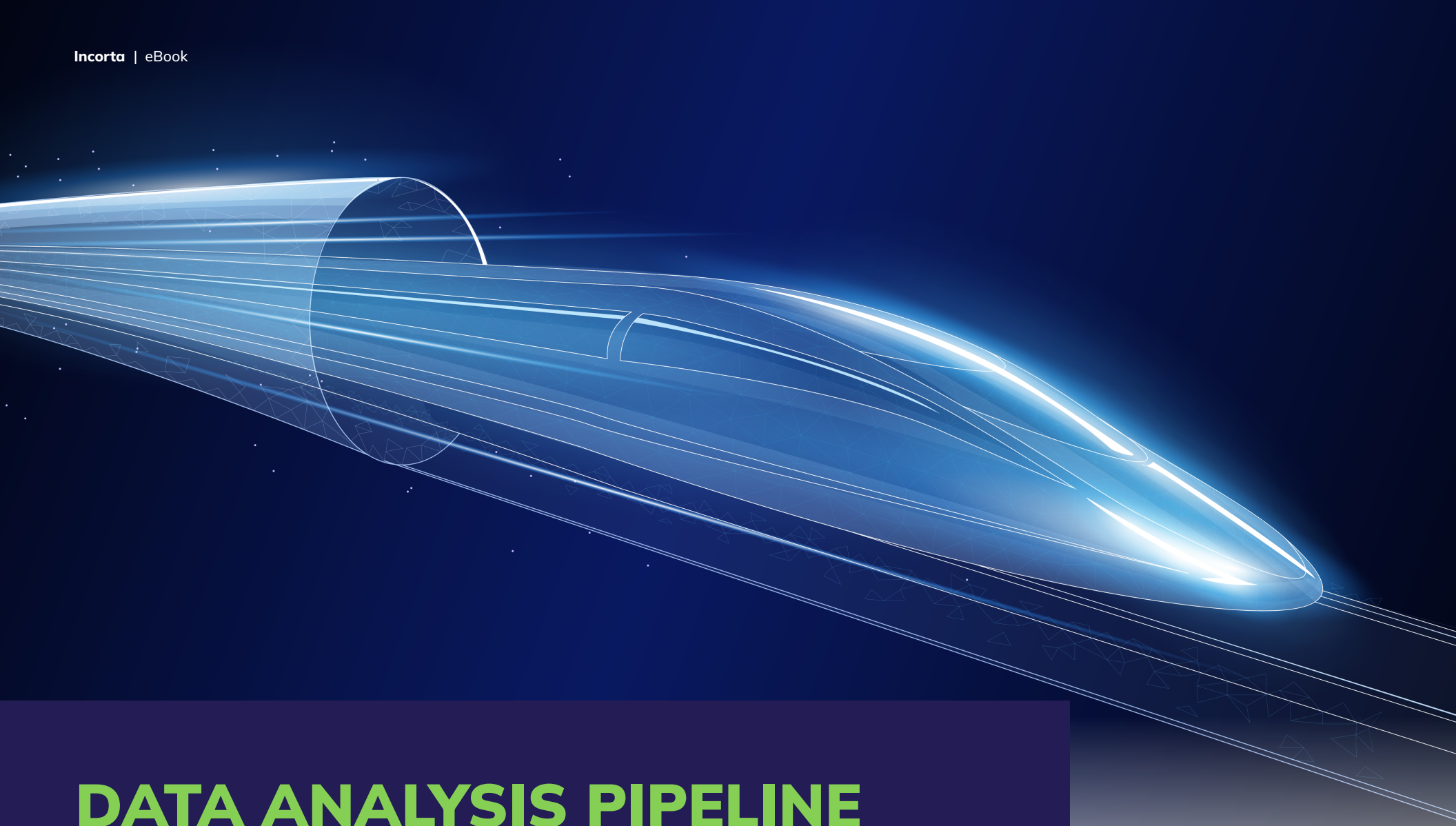


DATA ANALYSIS PIPELINE MODERNIZATION: STREAMLINING THE SLUGGISH MIDDLE

By: Greg Peters

incorta

According to estimates, 74 zettabytes of data will be created globally in 2021. That represents a nearly 20% annual compound growth rate since 2019.

No wonder organizations today are drowning in data. We all feel it … Data is streaming in from everywhere. It's coming from ERP systems, CRM and POS systems, supply chains — and that's just the start. To further complicate matters, the widespread adoption of IoT devices is causing an explosion in unstructured data. This segment of data is growing at a rate greater than 60% YOY.

In an attempt to manage this, organizations today have data routing to data lakes, data marts, warehouses, Excel spreadsheets, and BI tools. But it isn't working. There are too many data silos and data analysis pipelines are too slow, even despite the best efforts of technologists to speed it up over the past few decades.

With demand for real-time data analysis skyrocketing, it's time to take another look at data pipeline modernization — with fresh eyes and a fundamentally new approach.

So far, pipeline modernization efforts typically focused on two main areas: **(1)** the speed and capacity of data analysis pipelines, **(2)** the user experience. Now it's time to focus on streamlining and simplifying the middle part of the pipeline — the data transformation layer — or, **"the sluggish middle."**

Why haven't we seen more modernization at the data transformation layer? What happens when we do? To better understand how we got to this point, as well as the challenges and opportunities it presents for organizations of every size, let's take a stroll through history. As the saying goes, the best way to understand the future is to look to the past.

Rudimentary Pipelines

Back in the day, we used to write queries for data coming from the IBM AS/400 computer system. Introduced in 1988, it was considered revolutionary because it included hardware, an operating system, and a proprietary database.

There was a data analysis pipeline, albeit a very short one: data streamed out, and you could use it to make decisions. Unfortunately, AS/400 was a very difficult programming language, and the time it took to create a simple green bar report was measured in months, if not years.

Considering all the time and resources that went into it, it was commonly accepted that a report like this would cost a company as much as $100K. Even still, it was a big deal at the time, and considered to be worth the investment for some companies.

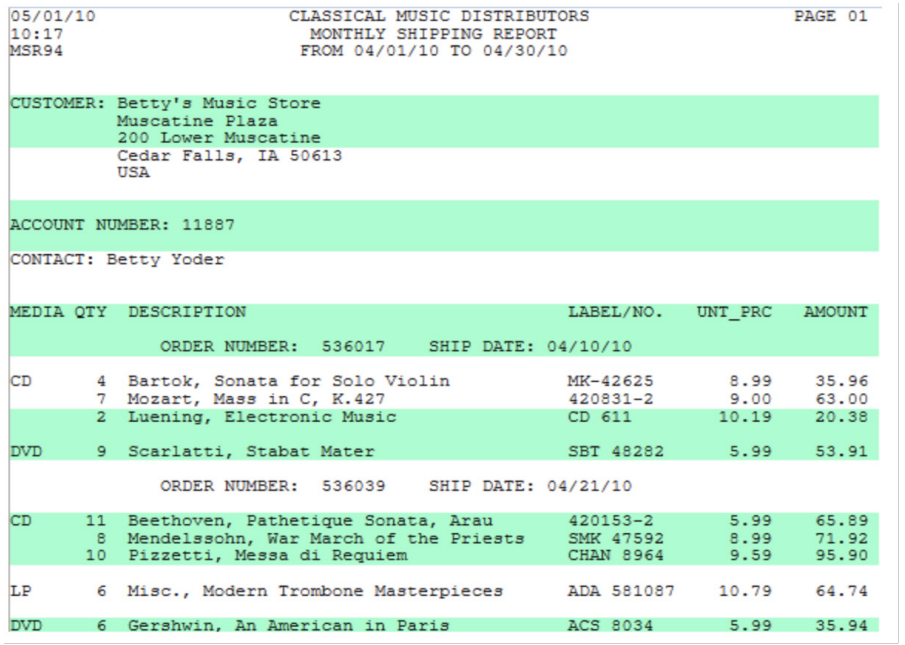
The Rise of Relational Databases

Data is the lifeblood of any well-oiled organization, and so it was only a matter of time before demand for data bumped up against the early limits of the technology.

Before long, we saw the rise of relational database players such as Oracle and Microsoft SQL Server, along with ETL vendors such as Informatica. These companies introduced new approaches for extracting data from an AS/400 database and loading it into a relational database for analysis.

This was an exciting moment. By making the barrier to data analytics much smaller, it meant that you no longer had to be an AS/400 programmer to do the analysis. Instead, you only had to know SQL, which was much easier to learn in comparison.

This protracted the solution stack, yet sped up the process by months. Most significantly, it began to democratize the accessibility of data. Now, a far broader range of people could work with data.



```
05/01/10                    CLASSICAL MUSIC DISTRIBUTORS                    PAGE 01
10:17                         MONTHLY SHIPPING REPORT
MSR94                       FROM 04/01/10 TO 04/30/10

CUSTOMER: Betty's Music Store
          Muscatine Plaza
          200 Lower Muscatine
          Cedar Falls, IA 50613
          USA

ACCOUNT NUMBER: 11887

CONTACT: Betty Yoder

MEDIA QTY  DESCRIPTION                              LABEL/NO.    UNT_PRC   AMOUNT

              ORDER NUMBER:  536017    SHIP DATE: 04/10/10

CD       4  Bartok, Sonata for Solo Violin          MK-42625       8.99    35.96
         7  Mozart, Mass in C, K.427                420831-2       9.00    63.00
         2  Luening, Electronic Music               CD 611        10.19    20.38

DVD      9  Scarlatti, Stabat Mater                 SBT 48282      5.99    53.91

              ORDER NUMBER:  536039    SHIP DATE: 04/21/10

CD      11  Beethoven, Pathetique Sonata, Arau      420153-2       5.99    65.89
         8  Mendelssohn, War March of the Priests   SMK 47592      8.99    71.92
        10  Pizzetti, Messa di Requiem              CHAN 8964      9.59    95.90

LP       6  Misc., Modern Trombone Masterpieces     ADA 581087    10.79    64.74

DVD      6  Gershwin, An American in Paris          ACS 8034       5.99    35.94
```

Drag-and-Drop BI

The next significant evolution came with the advent of "drag-and-drop" BI vendors, like BusinessObjects and MicroStrategy. This brought data analytics even closer into the hands of decision-makers.

All of a sudden, end users no longer needed to know SQL. They could simply drag and drop objects generated in the tool, blissfully insulated from the SQL working on the backend.

Despite these advancements, however, we were still stuck with the original challenge: If you wanted to analyze data, first you had to extract the data from an AS/400 or an ERP system, and then ETL it into an Oracle or SQL Server where it could be analyzed.

What's more, since the ultimate destination for the data was a business user, it also needed translating into a more digestible format. This added yet another layer into the mix — a metadata layer to translate the language of the database into the language of the business. This was sometimes within the BI tool, but not always.

Once again, the pipeline for accessing information grew even longer and more complex. We coped with it at the time because the benefits seemed to outweigh the costs: faster and easier access to data for analytics. In a way, it was like **trying to fill up a bucket with water while someone else randomly poked holes in the hose you were using to fill it up**. Meanwhile, the volume and velocity of water (i.e., data) that needed to go through the hose (i.e., data analysis pipeline) just kept growing and growing.

Massively Parallel Processing (MPP) and Hadoop

To address the exponential growth in volume and demand for data, the next phase of modernization aimed to increase the diameter of the pipeline itself. That is to say, we continued our quest for speed and accessibility – this time with more powerful appliances.

MPP systems such as Netezza and Teradata sped up Oracle databases and widespread adoption of the data warehousing methodologies of Ralph Kimball ratched it up even further.

With the dawn of the so-called "big data" era, Hadoop became very popular, largely because it was a free and completely open platform for storing all kinds of unstructured data. This helped firms handle the wider variety of data being generated by internet applications. In some instances, they could even replace Oracle and SQL dependencies with Hadoop and go much faster. End-user reporting that used to take 30 minutes or an hour could now be done in minutes.

In-Memory and Cloud

For all the allure of Hadoop, it ultimately became clear that it doesn't do so well with structured data from the data warehouse. At the same time this realization was happening, the price of RAM was plummeting. This gave rise to the development of very fast, in-memory databases such as SAP HANA.

At the time, machines to accommodate in-memory databases were still very expensive. This gave rise to yet another advancement: completely serverless cloud data warehouses such as Google BigQuery and Amazon Redshift.

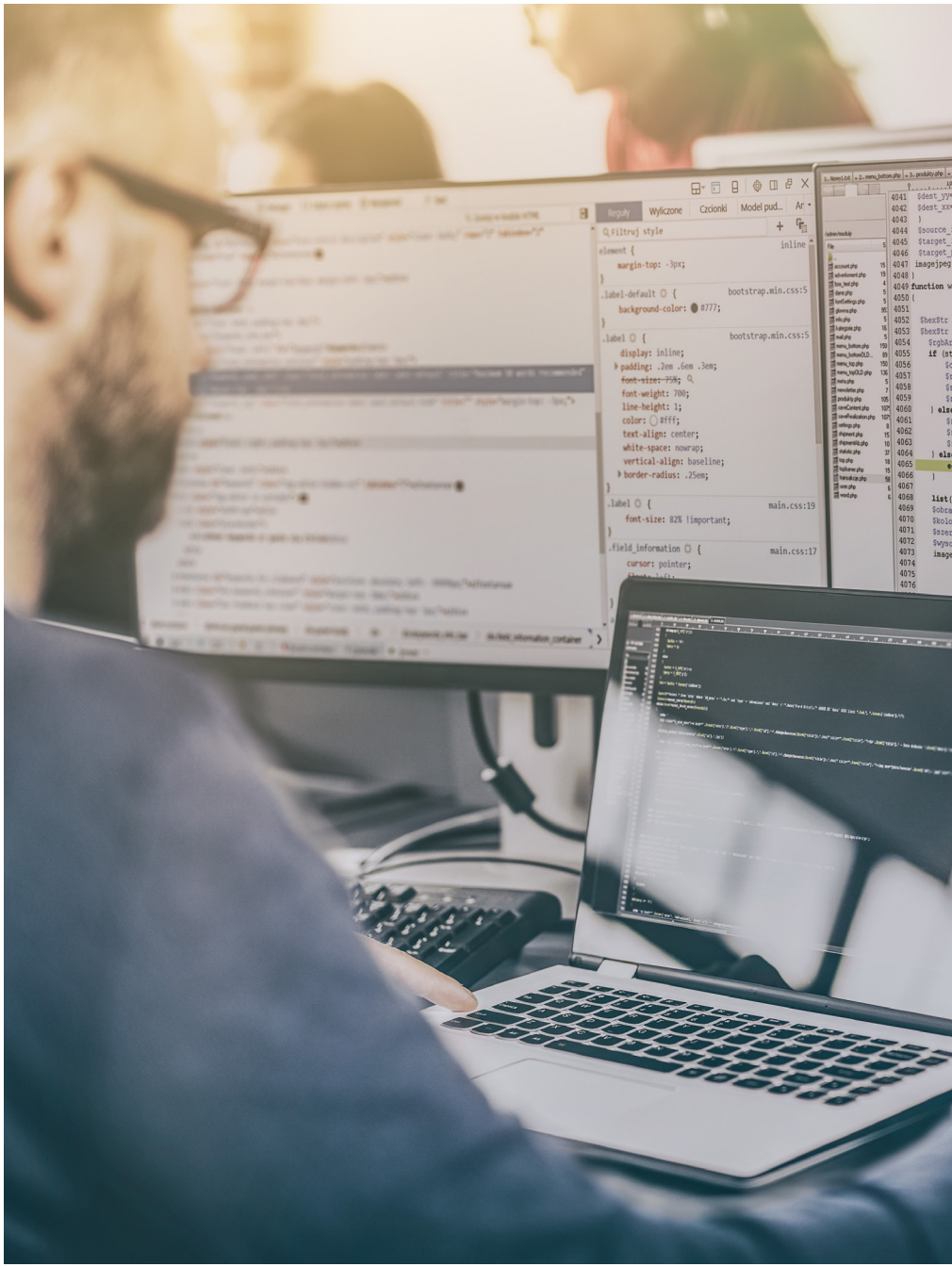
These new entrants promised all of the same speed and agility — and now you didn't even have to own and manage your own data warehouse. (To give credit where it's due, this advance did actually reduce some of the complexity for IT.)

Reducing Time to Insight

At this point in the story, we are well into the internet era, and business users are accustomed to working with data. They have new expectations, as well as new requirements for reporting and analysis — all of which create more demand for reporting tools, turning this once small subsegment into a proper market in its own right.

Meanwhile, BI is being modernized by players like Tableau and Looker (now a Google product). These tools allow people to visualize information on the fly, and they make those early BI systems look like band-aided report writers. The experience of visualizing information is now much richer.

We've come a long way from spending a year and $100K to pull a report out of an AS/400. But despite all of the technological advances, organizations still find themselves overly reliant on that sluggish middle layer – the *source-to-extract-to-transform-to-format-to-distribution* pipeline. In fact, even the most powerful BI tools available today still depend on this costly, burdensome framework.

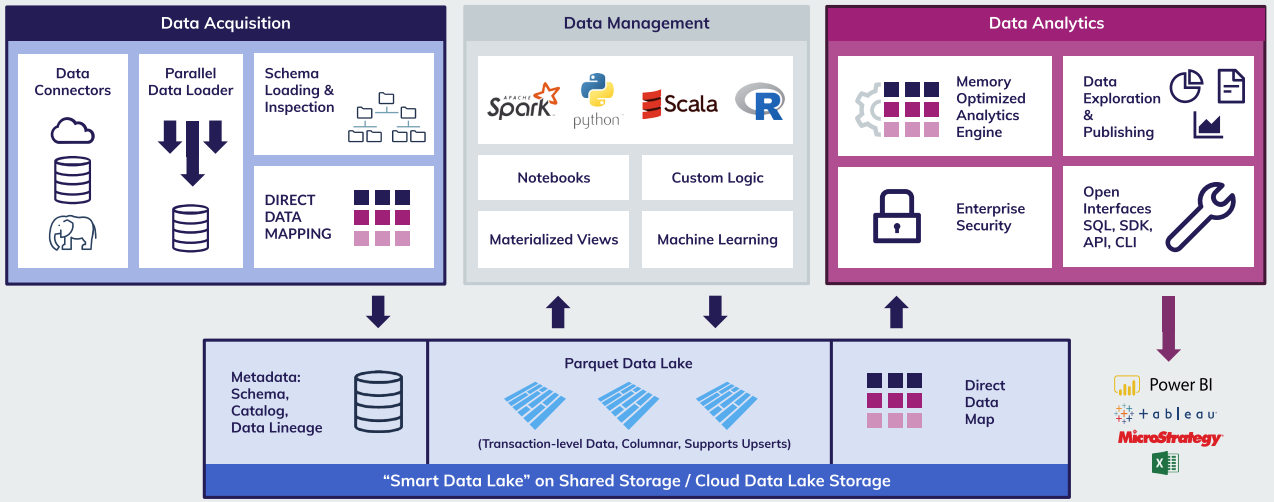


Streamlining the Sluggish Middle

The next phase of data pipeline modernization is all about streamlining the sluggish middle. After decades of workarounds, it's time to address the underlying problem — and now we finally have the technology to do it.

Incorta is a unified data analytics platform that introduces lightning-fast data analysis and visualization without the burden of ETL. By using Incorta, organizations sidestep the hassles of stitching together a patchwork of solutions and have a clear path to empowering individuals and teams with fast, efficient, and more effective analytics.

A simplified view of the Incorta unified data analytics platform



What happens when you streamline the sluggish middle with Incorta?

- You can connect to any data source without the need for slow and costly ETL processing and associated infrastructure.
- You enable self-service data access while maintaining robust security and data governance controls.
- You help data scientists and analysts become more productive by enabling them with rich, intuitive interfaces that make it easy to realize valuable insights from data.
- You are no longer forced into costly data warehouse infrastructure. Instead, you can realize the benefits of a simple, powerful unified data analytics platform at your own pace, gradually reducing your reliance on legacy systems as it suits you.
- You start making fast progress, deploying in a matter of minutes as a fully managed service that is open and extensible, and supports deployments on premises or on a customer's preferred public or private clouds.

For a more detailed look at Incorta's architecture and how various aspects of our unified data analytics platform works, check out the Incorta Architecture Guide for a deep dive.

For those who prefer a more hands-on approach, start a free trial and see Incorta in action for yourself.

THE DIRECT DATA PLATFORM™

ABOUT INCORTA

Incorta is the data analytics company on a mission to help data-driven enterprises be more agile and competitive by resolving their most complex data analytics challenges. Incorta's Direct Data Platform gives enterprises the means to acquire, enrich, analyze and act on their business data with unmatched speed, simplicity and insight. Backed by GV (formerly Google Ventures), Kleiner Perkins, M12 (formerly Microsoft Ventures), Telstra Ventures, and Sorenson Capital, Incorta powers analytics for some of the most valuable brands and organizations in the world. For today's most complex data and analytics challenges, Incorta partners with Fortune 5 to Global 2000 customers such as Broadcom, Vitamix, Equinix, and Credit Suisse. For more information, visit https://www.incorta.com