incorta

FIVE CONSIDERATIONS FOR BUILDING AN AGILE DATA PIPELINE IN THE CLOUD

Introduction

In today's hyper-competitive business environment, data is at the heart of everything we do. Business leaders rely on up-to-date data to better understand their customers and competitors, make better, more informed decisions, and support new business initiatives.

Unfortunately, the traditional ways of collecting, curating, and analyzing data no longer meet the needs of most organizations. The basic architecture of the data pipelines that collect, transform, and land data in enterprise data warehouses and data marts have hardly changed in decades. They are often slow, inflexible, and costly to operate and maintain.

Decision-makers can no longer wait weeks or months for information. Speed, agility, and access to the latest information have become a matter of survival. Fortunately, innovations such as smart data connectors, cloud-friendly file formats, and new in-memory analytic techniques can dramatically streamline data pipelines and, in some cases, eliminate them entirely.

This paper will examine some of the challenges with existing data pipelines and discuss five considerations for building a more agile data pipeline infrastructure in the cloud. It will then introduce Incorta, a modern unified data analytics platform (UDAP), and explain how Incorta can help organizations increase the effectiveness of their data and analytic environments.

A Challenging Data Landscape

Transactions have increasingly shifted online, fueled by megatrends such as the internet, mobile, and cloud computing. With plummeting collection and storage costs, and new techniques for analyzing a wider range of data, organizations are collecting and retaining more data than ever before. According to Statista, worldwide data storage capacity is expected to grow at a CAGR of 19.2% through 2025.[1] Despite being awash in data, most organizations struggle to turn their raw data into useful business insights. In a recent survey of 215 analytics and business intelligence (BI) decision-makers, 90% of respondents indicated that their current solutions could not meet all their business objectives.[2] Despite throwing millions of dollars at their data infrastructure, organizations often experience frustration and limited returns.

1 Statista – Total data volume worldwide 2010-2025

2 Forrester Consulting, June 2021 – Accelerate Business Insights With a Modern Full-Stack Analytics Platform

Smart business leaders know that innovation occurs when talented people with firsthand knowledge of the business are given the tools they need to explore and experiment with data. However, delivering this kind of self-service environment is surprisingly difficult in practice.

Data engineers are tasked with stewarding the business's data assets and information resources. They face a variety of challenges, as illustrated in Figure 1. These include fast-growing datasets, increasing data access requests, and often stale data of questionable lineage and accuracy. They are also constrained by limited IT budgets and growing challenges related to security, governance, and compliance.

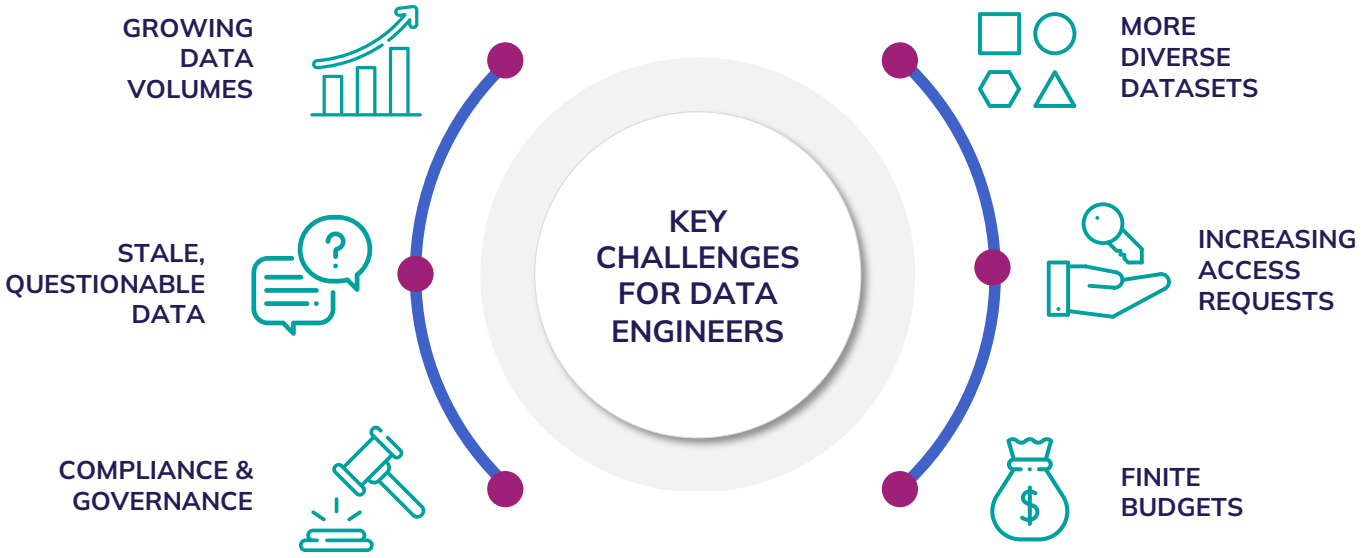

Figure 1 – Data engineers face a variety of challenges

While the data landscape has changed dramatically, solutions for collecting and managing data have not kept pace. The basic process of extracting data from operational systems, transforming it, and loading it in a data warehouse (referred to as "ETL") has remained largely the same for more than 30 years.

Traditional Approaches Fall Short

To understand why data environments remain so challenging, it is helpful to examine how they are constructed today. In most organizations, data resides in various operational systems. Large enterprises may operate a dozen or more systems for everything from manufacturing to finance to HR. Examples include systems for enterprise resource planning (ERP), supply chain management (SCM), customer relationship management (CRM), and more.

As customer and partner transactions have shifted online, organizations have seen value in collecting and analyzing additional types of data as well. These nontraditional sources include web logs (tracking customer clickstream activity on the corporate website), telemetry from mobile apps, API logs, call center interactions, social media activity, etc. This data is frequently landed in a data lake in a raw form.

By collecting data from these diverse sources, organizations should be better positioned to distill all this data into useful information. The challenge is that data in its raw form is poorly suited to data exploration and analysis. Operational databases are optimized to process transactions efficiently and maintain data integrity. These databases are typically organized in third normal form (3NF), with complex schemas that spread data across

hundreds or thousands of tables in a relational database.[3] Even if it were possible to query these databases directly, most customers would not do it for fear of interfering with critical business operations.

To make operational data accessible for reporting and analysis, most organizations extract and reshape data into a form that is easy to query and analyze. A high-level illustration of this process is provided in Figure 2. Data is transformed and refined, gathered from upstream sources, and landed in a data warehouse. Business-friendly data views are often created in topic-specific data marts.

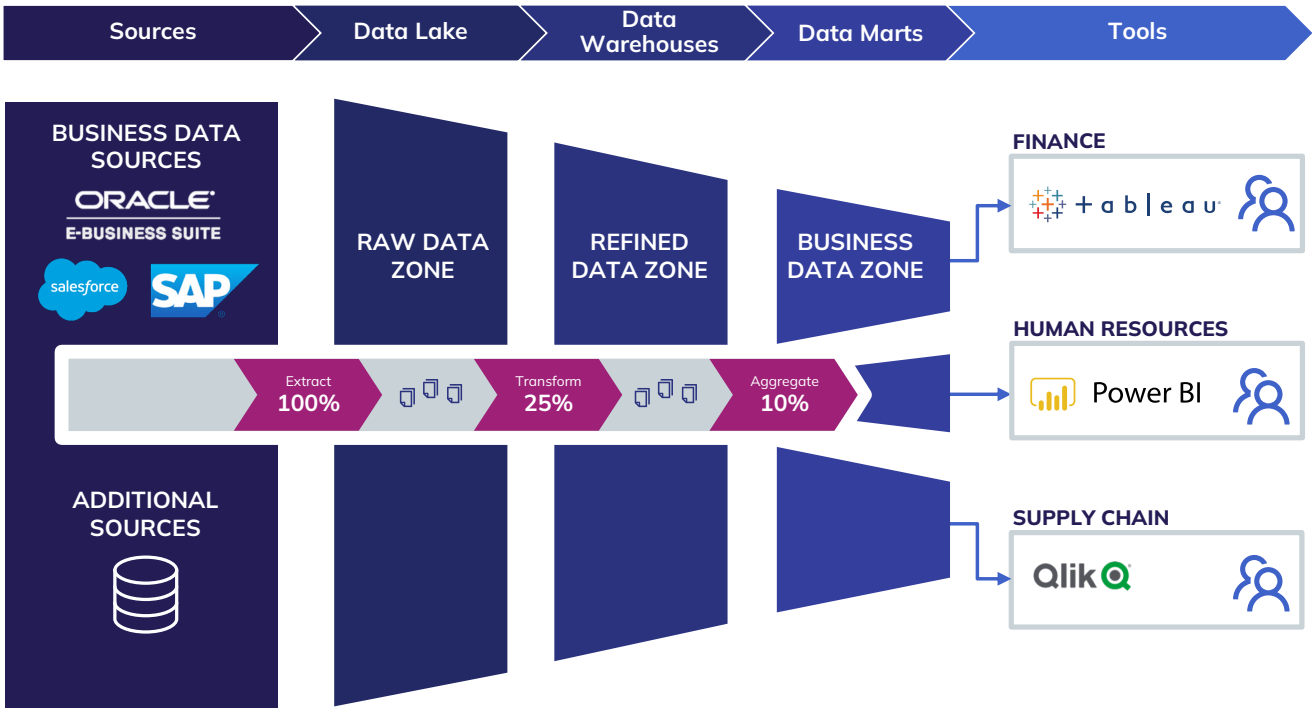

Figure 2 – The traditional approach to business reporting and analytics

Data warehouses typically store data using a star or snowflake schema, where business facts are surrounded by various dimensional tables pre-aggregated to support multidimensional analysis.[4] Business users and analysts then use BI tools such as Tableau, Power BI, Qlik, and others to generate reports and view business-friendly dashboards.

Data engineers spend considerable time and effort to realize the data architecture illustrated in Figure 2. They carefully design the star schemas required to support reporting and decision support requirements efficiently. They also need to build and maintain automated data pipelines that refine data and make it accessible to business users and analysts.

3 Statista – Third normal form – Wikipedia

4 A star schema is an approach widely used to develop data warehouses and dimensional data marts. Star schema – Wikipedia

Data Pipelines Add Cost and Complexity

Data pipelines lie at the heart of traditional data management environments. They manage the process by which data is extracted from operational databases, cleansed, aggregated, and reshaped into a form that can be easily queried. Pipelines are used for many purposes, including moving data to and from data lakes, preparing data extracts and OLAP cubes, periodically purging or archiving stale data, and preparing datasets to train predictive models.

Regardless of where they are used, there are several challenges associated with data pipelines:

- Pipelines are expensive, resource-intensive, and difficult to manage
- They can be complex, brittle, and difficult to maintain
- They require ongoing monitoring and management
- They can be slow and run only periodically, leading to old, stale data
- Every time data is filtered or aggregated, we get further from the ground truth

Agility Is Key to Business Competitiveness

Today the business landscape is more competitive than ever. The global pandemic has demonstrated that being agile is no longer just a business differentiator. It is an essential capability for organizations to prevail against external threats such as supply chain disruptions, fluctuating commodity prices, changing labor markets, and climate change.

To thrive, organizations need to innovate quickly to satisfy customers' changing needs, and swiftly adapt to market shifts. In business terms, agile organizations:

- Anticipate customer needs and quickly respond with new offerings that boost revenue and profits
- Quickly adopt new technologies such as machine learning and AI-based services to boost efficiency
- Easily adapt to changes in the regulatory landscape or reporting requirements
- Engage in mergers and acquisitions, easily shift priorities, and quickly refocus the business
- Refresh data intraday, and close the books on time

Unfortunately, existing information systems frequently get in the way of these goals. Slow and inefficient data pipelines make it challenging to incorporate new data sources, obtain accurate and up-to-date information, and achieve a clear line of sight to business operations.

Five Considerations for Building an Agile Data Pipeline

In their current state, data pipelines represent a challenge to organizations attempting to innovate quickly and deploy new services. As organizations struggle to build more agile, responsible pipelines, here are some important considerations to keep in mind:

1. Friends Don't Let Friends Transform Data

One of the challenges with transforming data is that every transformation takes us further from the ground truth. During ETL processing, organizations discard some data fields and aggregate others, summing, averaging, and grouping records. The problem with this is that when anomalies occur, analysts lack an easy way to drill down to the source and find an underlying cause. The more complex the pipelines, the more removed decision-makers become from the source.

A second challenge is that every transformation creates debt — in both a business and technical sense. Each new transformation requires a cottage industry to support it: ETL developers, DBAs, BI developers, and dashboards artisans. The more moving parts in a pipeline, the more failure-prone it becomes. Changes in data volumes, source schemas, or transient server or network outages can cause pipelines to fail. These failures lead to organizations being unable to refresh data intraday, missed deadlines, lack of visibility to the business, and associated opportunity costs.

Business users are the experts in their own domains. However, requests for new data views can take weeks or months due to data engineering backlogs. Deriving new data views typically require new pipelines, modifications to existing pipelines, or changes to data schemas. Pipeline changes need to be carefully planned, implemented, and thoroughly tested before being put into production. It is hard to spot and react quickly to a market shift when it can take months to gain access to the data.

To combat these challenges, organizations need to rethink their data architectures. They need solutions that avoid costly data transformations. Ideally, business users should be able to access and analyze data they are entitled to see without assistance from data engineers. Pipelines should be responsive and accommodate new data sources instantly without operational disruption.

2. Performance Matters — From Multiple Perspectives

Performance is at the heart of most BI and analytic challenges. In fact, performance is why we transform and load data into downstream data warehouses in the first place. As explained earlier, running analytic queries against transactional databases is impractical given how data is structured. Queries would be horrendously complex, involving complex joins across multiple tables and data sources. The whole point of reshaping data is to support fast, efficient multidimensional analysis. Query speed is essential, but there are other performance-related considerations as well:

- How quickly can we load/synchronize data from an external source?
- How frequently can data views be refreshed?
- How long does it take to make a new data source accessible to an analyst or business user?
- How long does it take for a pipeline to reflect a change in source or target schemas?

These aspects of performance help drive what matters to the business — better organizational agility and faster time to insight. Whether people are analyzing top-line KPIs or transaction-level detail, easy access to the latest data can change an organization's culture. An organization with self-serve access to data is more confident that answers to new questions can be obtained quickly and accurately. Users are empowered to be more forward-looking, ask questions, and challenge existing notions of doing things. Organizations empowered with fast access to up-to-date data are more innovative, productive, and profitable.

3. Don't Accept Trade-Offs Between Self-Service and Data Security

Users demand self-service access to data, but the data security and governance landscape is becoming more complicated. Cyberthreats, data leaks, and compliance are all top-of-mind concerns in corporate board rooms. Managers may perceive security threats to operational systems as the most significant risk, but analytic and reporting systems are vulnerable too. For example, the well-publicized Colonial Pipeline ransomware attack in 2021 never actually reached control systems. Rather, it involved compromised credentials for a downstream billing application.[5] The potential financial and reputational damage from these types of threats is enormous.

Traditionally, to provide business users with better access to data, data engineers have sometimes compromised, providing access to data without proper security and governance controls. Examples include creating flattened views for reporting or downloading disconnected copies outside the governance purview of the data warehouse. Users often turn to Excel as a tool of last resort, and warehouse administrators lose control over sensitive data entirely.

In today's threat environment, the risks are simply too great. This is especially true in healthcare institutions, banks, and organizations subject to legislation around the handling of personally identifiable information (PII).[6] Fortunately, with modern tools, enabling self-service data access does not mean compromising security. Solutions exist to securely provide business users with self-serve data access without significant data engineering effort.

4. Stay Open, Portable, and Flexible

Data sources, schemas, and analytic tools are constantly evolving. Understandably, organizations are looking for new cloud-based solutions to help them manage data more effectively. Unfortunately, there are multiple ways that organizations can find themselves locked into specific data architectures and vendors. Some examples include:

- Building a data management environment tied to a cloud provider's proprietary data warehouse
- Using tools that work with only a single data source or warehouse offering
- Using tools that limit the data sources that you can access

Today's business environments are dynamic. Large organizations invariably use multiple tools and have multiple data sources. This occurs because of mergers, acquisitions, or different groups making independent purchasing decisions.

While cloud data warehouses can be convenient, organizations can find themselves constrained in their choice of tools and burdened with hidden costs or opaque pricing structures. For example, a cloud provider's data warehouse may allow users to read data from an open data lake file or table format. However, they may charge per GB read — a practice that can get costly fast depending on data access patterns and where data is stored.

Organizations need flexible solutions that do not lock them into data sources, warehouse providers, or cloud environments. At the same time, however, they need to avoid the hassles and expense of managing dozens of discrete tools from different vendors stitched together to process a data pipeline. Data environments must be instantly adaptable to changing business requirements and support connections to all data sources and destinations. The flip side of being locked-in is being locked-out of opportunities for change. Staying open makes you ready for tomorrow's challenges, whether moving to the cloud, working with new data types to support key initiatives, or building bridges between departments and technologies.

5 Colonial Pipeline ransomware attack – Wikipedia

6 Examples include the EU General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and Canada's PIPEDA (Personal Information Protection and Electronics Document Act).

5. Don't Reinvent Wheels

According to Forrester Consulting, 69% of organizations say that data preparation is one of the most time-consuming aspects of analyzing data.[7] Nevertheless, organizations frequently spend time independently solving the same problems repeatedly. Each organization builds and maintains data pipelines to extract and reshape data from off-the-shelf ERP systems and cloud SaaS platforms. Industry-wide, this represents a massive waste of resources. Platforms such as SAP, Oracle, and Salesforce have complex schemas comprising thousands of tables. Organizations can spend years understanding these schemas, modeling data, and devising data views useful to business users and analysts.

Ideally, organizations should look for solutions that expose datasets and dashboards in a form where they are instantly usable. They should also look for platforms that make it easy to combine data across multiple upstream business applications.

Challenging Existing Thinking About Data Pipelines

Current approaches to managing and analyzing data often hinder the organizations' ability to meet business imperatives. Data pipelines introduce complexity, latency, and impede their ability to innovate and differentiate themselves from competitors. Pipelines are often comprised of a patchwork of legacy tools not designed to handle the volumes of data and inherent complexity of modern data analytic problems.

Enabled by new technologies, a new category of data analytics platform has emerged that challenges traditional thinking about data pipelines and analytics. Long-held notions considered to be "set in stone" such as the need for ETL pipelines, star schemas, data cubes, and data marts, are being upended.

By freeing themselves from these traditional constraints, organizations can benefit substantially — reducing cost, complexity, and improving the speed and efficiency of their analytic operations — enabling users to be more innovative and dramatically accelerating time to insight.

7 Forrester Consulting – Accelerate Business Insights With a Modern Full Stack Analytics Platform

Incorta — A Better Approach

Incorta is a unified data analytics platform (UDAP) that delivers high-performance data analytics without the cost and complexity of traditional solutions. It is a complete end-to-end data platform that integrates data acquisition, storage, analysis, and visualization into a single platform. A high-level view of major Incorta components is provided in Figure 3.

At the heart of Incorta is its Direct Data Mapping™ (DDM) technology that delivers fast query performance without the need to manually reshape or transfer data. This means that users can access and analyze source data directly. It also means that data engineers can spend less time building and maintaining complex pipelines and separate data warehouses to support business users and analysts.

Incorta combines an open data lake with in-memory analytics, as illustrated in Figure 3, for both scalable storage and responsive queries. This architecture supports both durable storage and responsive queries. Incorta serves multiple workloads, including big data processing, data enrichment, and machine learning via a built-in Spark environment with data science notebooks. This translates into less friction, more consistent answers, and faster innovation.

Business analysts can opt to use Incorta Analyzer to create dynamic, intuitive data dashboards within the UI. Alternatively, they can access business-friendly data views from existing tools, including Tableau, Power BI, and Looker, through a PostgreSQL compatible open interface.

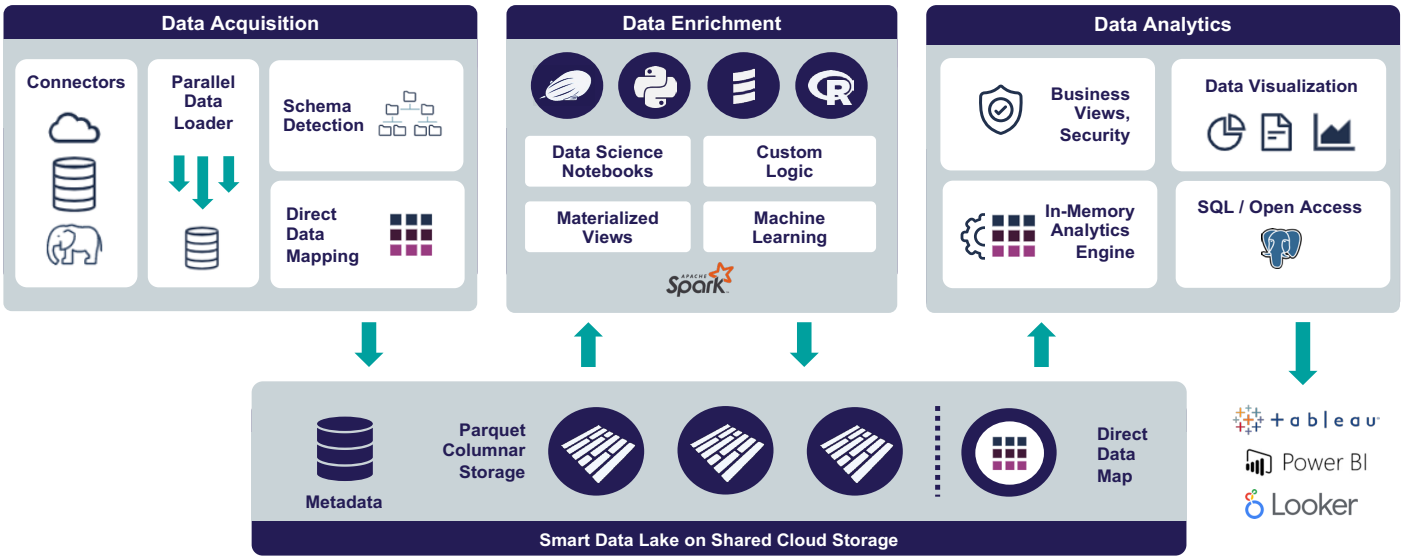

Figure 3 – Incorta high-level architecture

Unlike traditional analytic environments that perform ETL and store summary data in carefully designed star schemas for analysis, Incorta performs "ELT" — quickly ingesting data from on-premises or cloud-based sources and storing it in a smart data lake. Data resides on shared cloud storage in modern, open Parquet file formats. Not only does Incorta avoid the overhead of slow and expensive transformations, but data is preserved in its original form, providing analysts with complete visibility to their data for better decision making.

Intelligent Data Ingest

One of the keys to Incorta's exceptional ease of use is its innovative Data Loader service. The Loader Service is responsible for extracting data from external data sources via connectors and persisting physical schemas in shared storage. Incorta can connect to virtually any database, enterprise application, data stream, or file format with 240+ connectors provided by Incorta and Incorta partners.

The Loader Service can either load data on demand or use a pre-built scheduler to load data into physical schemas at user-configurable intervals. Parallel loading and data chunking deliver exceptional performance to ensure that analysts always have access to the most up-to-date data. Connectors can employ different incremental loading strategies rather than performing a complete database extract, periodically fetching only data that has changed since the last successful extract.

Making ETL History

Incorta's DDM technology automatically analyzes source schemas as data is ingested, determining which columns to treat as "measures" and "dimensions" for analysis purposes. Incorta considers column names, data types, cardinality, and relationships among data sources when making these determinations. By pre-processing normalized data sources in a fashion that anticipates all potential query paths and storing data in query-optimized Parquet and DDM files, Incorta supports lightning-fast queries.

For a business analyst, this means that you can query source data directly while realizing the performance of data that has been pre-processed or flattened, but without the need for ETL processing. From an IT perspective, users can have access to analytics-ready data, but without the time and resource-intensive data preparation steps required by traditional approaches.

In traditional warehouse environments, obtaining access to new data for analysis was tedious and time-consuming, as illustrated in Figure 4. To obtain data required to answer a business question, analysts often need to contact their IT organization and obtain assistance from a data engineer. Building and validating new ETL workflows and creating new schemas for the data warehouse or data mart could take weeks or months. Even when new pipelines and data views were created, analysts still only had access to summary aggregates. This impeded their ability to understand data and make effective decisions fully.

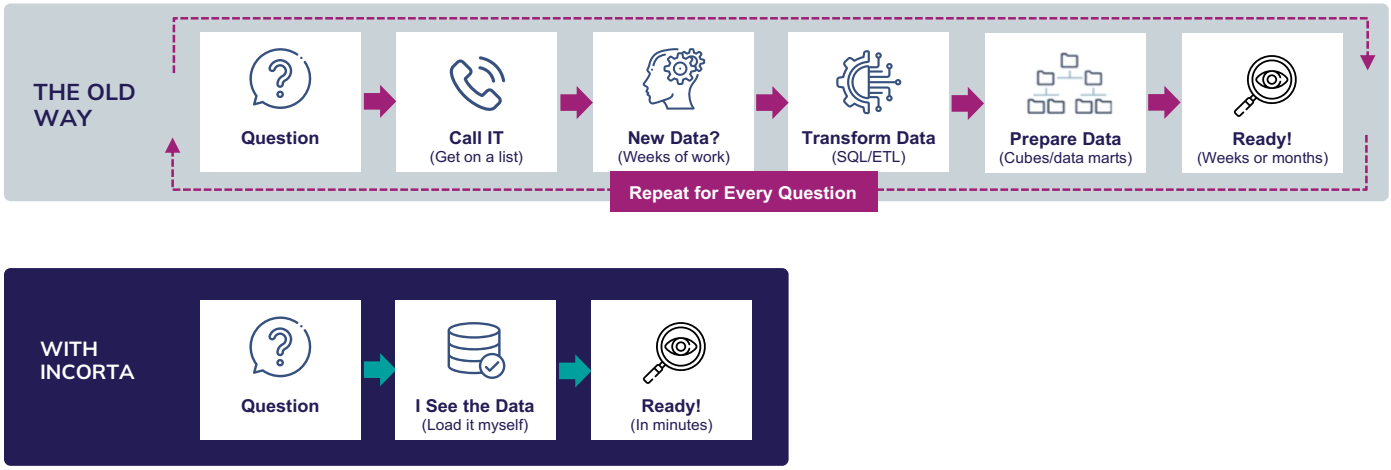

Figure 4 – Incorta helps dramatically reduce time to insight

Incorta virtually eliminates the need for complex ETL and data reshaping. Data from any external source can be ingested using an automated Schema Wizard. Once ingested, analysts and business users have immediate visibility to source data subject to Incorta security and data governance controls. Users can see data immediately, trigger updates whenever they want, and easily create derived views, dashboards, and visualizations, often within minutes.

Incorta Blueprints

A key feature of Incorta is its collection of business-friendly Blueprints. Blueprints provide pre-built schemas and dashboards for accessing, organizing, and presenting data from popular business solutions based on best practices. Blueprints are provided for multiple applications, including Guidewire, NetSuite, Oracle ERP Cloud, SAP, JD Edwards, Oracle EBS, and Salesforce as shown in Figure 5. Business schemas address multiple functional areas in each application and pull together key metrics, sample reports, visualizations, and sample dashboards. Examples include dashboards for accounts payable (AP), accounts receivable (AR), bill of material analytics, fixed asset analytics, enterprise asset management, order fulfillment, procurement and spend analytics, etc.

For enterprise users, Blueprints are an enormous timesaver. Analysts and data engineers no longer need to spend time understanding the complex schemas associated with these enterprise applications, building pipelines, and creating tables in a data warehouse. Blueprints do the heavy lifting, automatically connecting to source tables and reflecting business views and dashboards in Incorta so that customers typically require only light customization after installation.

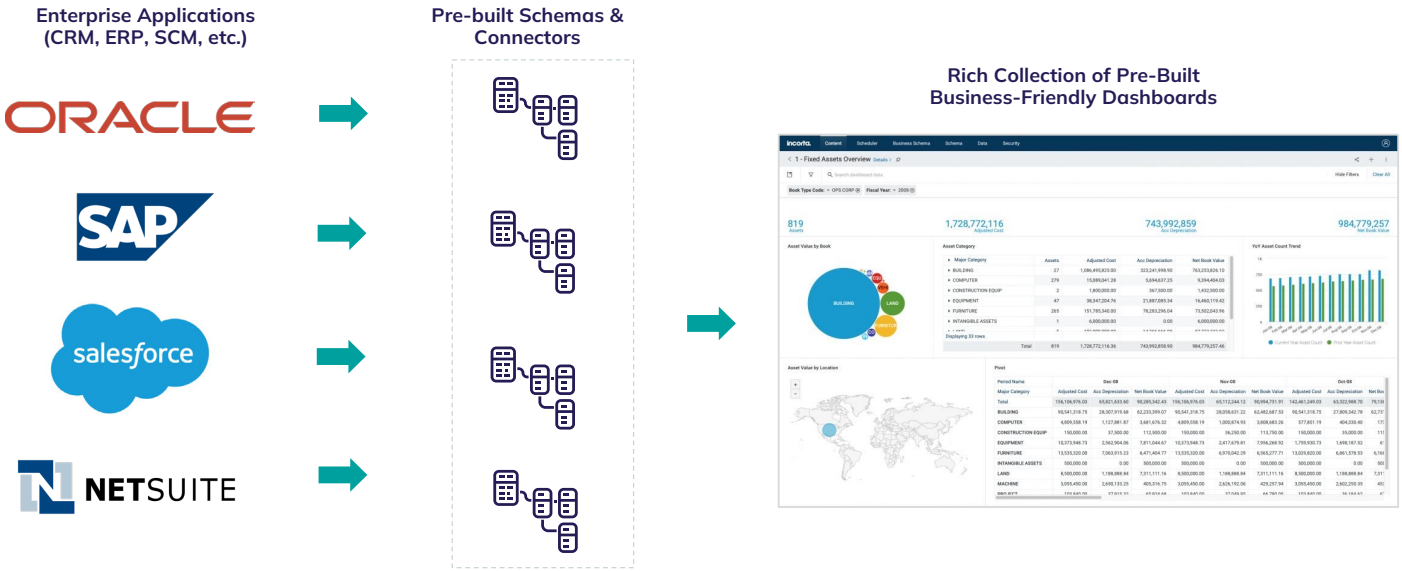

Figure 5 – Incorta Blueprints help organizations get up and running quickly with enterprise applications

Delivering Agile Cloud Data Pipelines

For organizations seeking to realize a more agile analytics environment, Incorta provides unique value. It can often eliminate the need for expensive data warehouses, data marts, and data integration solutions and dramatically improve the performance and agility of cloud data pipelines.

Incorta directly addresses the challenges with traditional pipelines discussed earlier, helping organizations be more agile and respond to changing business requirements. Unique capabilities of Incorta are:

- **No ETL required** – Incorta virtually eliminates the need for costly and complex ETL workflows and associated data engineering. New data sources can be added in minutes and exposed in business-friendly formats with no loss in fidelity.
- **Exceptional performance** – Analysts enjoy better performance all around. With Incorta's efficient Loader Service, decision-makers always have access to the latest data.
- **Secure, self-serve access** – Rather than waiting for data engineers to create new data views, business users and analysts enjoy self-serve access to data without compromising on security or data governance controls.
- **Open, portable, and flexible** – While Incorta provides rich functionality, it fully protects existing investments by storing data in open, cloud-friendly file formats. Customers can deploy Incorta in their environment, connect with any data source, and access Incorta data from any BI tools.
- **Incorta Blueprints** – Finally, rather than reinventing wheels and spending time and effort building warehouse schemas and dashboards for enterprise applications, customers can leverage Incorta Blueprints, providing instant access to business-friendly data views without reinventing the wheel.

Incorta works seamlessly with existing on-premises or cloud-based warehouses and data integration tools, allowing customers to gradually simplify their environments and improve the quality of their analytics at their own pace.

Summary

Today, access to accurate and up-to-date information is essential for business competitiveness. Organizations rely on information to understand their customers and make better, more informed decisions.

Unfortunately, the traditional way of collecting, transforming, storing, and analyzing data can no longer keep pace with changing business requirements. Existing data pipelines are often slow, inflexible, and costly to operate and maintain.

By keeping in mind some simple guidelines, organizations can modernize their pipelines and create an information ecosystem that is agile and responsive. As a cloud-based alternative to traditional pipeline infrastructure, Incorta can provide an essential role in this transformation.

By using Incorta, organizations avoid the hassles associated with managing and stitching together multiple discrete solutions and can empower the business with fast, efficient, more effective analytics. Organizations can gradually reduce their reliance on costly data warehouse infrastructure and realize the benefits of a simple, powerful unified data analytics platform at their own pace.

To get started with a free trial of Incorta Cloud, and jump-start your analytics, visit cloud.incorta.com/signup.