INCORTA ARCHITECTURE GUIDE

November 2021

Incorta's revolutionary analytics technology sidesteps the traditional data workflow to enable easy user adoption, rapid insights, and better business results.

incorta

Introduction

Now, more than ever, information is a competitive weapon. Businesses literally win or lose based on the effectiveness of their data analysis. Good data impacts everything from improving operational efficiency to creating better customer experiences to strategically expanding product offerings and markets.

However, the traditional ways of collecting, curating, and analyzing data are complex, inflexible, and costly. Top-down business intelligence initiatives, with sweeping objectives including canonical models, extensive data pre-processing, and spoon-fed reports, were never ideal, and today are wholly inadequate.

Business leaders require, and are demanding, a new data paradigm — agility in the face of fresh challenges — a way to achieve timely analysis and fast execution.

This guide provides a detailed description of Incorta, a modern unified data analytics platform that avoids the pitfalls of traditional data warehouses and analytic systems. We examine Incorta's architecture and describe how it can augment or replace legacy analytics environments while delivering new capabilities to boost productivity and competitiveness.

Contents

The Case for Incorta

We are at a data crossroads in the business world. As the pace of change increases and new challenges mount, data-driven decision-making and data experimentation are more important than ever before.

Unfortunately, most existing analytics platforms are not architected to meet today's demanding business environment, where information and insights are required in real time. In a recent survey of 215 analytics and business intelligence decision-makers, 90% of respondents indicated that their current solutions could not meet all of their business objectives.[1]

The Rise of the Citizen Analyst

Organizations have been trying to efficiently turn data into business insight for decades. New tools, technologies and platforms have promised, and failed to deliver on ubiquitous, data-informed decisions. Companies continue to throw millions of dollars at this problem with limited, frustrating returns.

Worse, long lead times to deliver on data and analytics initiatives tend to stifle initiative and enthusiasm for asking the important questions, iterating and drilling down on those questions, and driving innovation.

In today's business environment, data leaders can no longer wait weeks or months. Speed, agility, and timely access to the latest data has become a matter of survival.

At the same time, smart managers realize that innovation occurs when ambitious people are given the tools and freedom to explore and experiment with business data.

Self-service analytics is on the rise because teams are getting access to more intuitive data analysis tools. However, it isn't only tools that lead to data insights. Business professionals need to be data literate and they need to understand what questions to ask to get meaningful insights.

Business-led analytics initiatives that enable users to build analytic models and dashboards over time in an iterative, agile manner, are creating a new culture of data-driven decision-making.

With the proper education and access to tools, business professionals can do a better job of analyzing their own data given that they have firsthand knowledge of their line of business. Removing the intermediary speeds up time to action and the ability to pivot more quickly.

---

[1] Source Forrester Consulting, June 2021 – Accelerate Business Insights With a Modern Full-Stack Analytics Platform

New Business Imperatives vs. Technical Constraints

Data analytics, the ability to correlate multiple variables together in an ad hoc fashion (as opposed to static reporting or business intelligence), has become an essential competitive tool. Faced with an increasing pace of change, organizations need to anticipate and quickly adapt to unexpected market changes and moves from competitors.

Business leaders face a gauntlet of critical imperatives:

- *Accelerate time to value* – by embracing agility and avoiding top-down projects with large scopes, even larger budgets, and no guarantee of success.
- *Enable critical new applications* – by leveraging a modern data analytics pipeline to reduce time to market for new strategic applications and initiatives.
- *Democratize data access* – by providing access to data, tools, and the freedom to explore and experiment, unleashing natural innovation.
- *Comply with new regulations* – by handling personally identifiable information (PII) in a way that doesn't compromise the ability to ask questions.

Administrators of data and analytic systems also face an increasingly challenging data environment and IT landscape:

- *Growing volumes of increasingly diverse data* – making it challenging for organizations to collect, curate, understand, and analyze all of their data.
- *More requests for access to new and existing data sources* – resulting in costly delays and wait times as data engineering requests mount. Additional changes place an added strain on already complex Extract-Transform-Load (ETL) and Extract-Load-Transform (ELT) workflows and tight batch windows.
- *Multiple data stores, stale datasets, and data of questionable lineage and accuracy* – complicating data governance and making it difficult to obtain timely, high-quality insights.
- *Finite IT budgets and a tight labor market* – making it increasingly difficult to deal with the opportunities and challenges described above.

Business and IT teams stand at the nexus of this data integration and management challenge. They need to enable citizen analysts and ensure that business imperatives are met, and protect the organization from a variety of dangers including the inability to scale. This dilemma is illustrated in Figure 1.

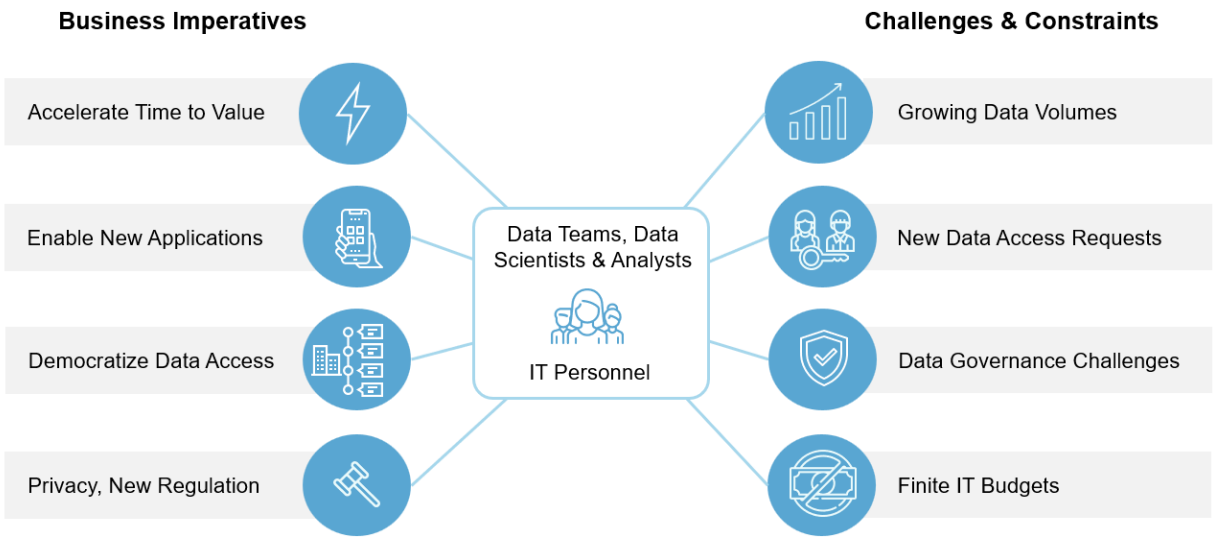

Figure 1 – Business and IT teams face opposing imperatives and constraints

Existing Solutions Fall Short

Presently, most organizations operate multiple tools and data management platforms. These include data warehouses, data lakes, data integration and ETL frameworks, and various data science and business intelligence (BI) applications for decision support.

The data warehouse is typically fed by multiple upstream data sources, including Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM) systems, transactional databases, and other sources. The volume and variety of these data sources are a growing concern. The same Forrester Consulting study referenced earlier found that enterprises draw data from an average of **320 different data sources** to feed downstream analytic and BI applications. This number is **expected to grow to 708** in just the next 12 months — a staggering **121% increase** in the number of data sources.[2]

Organizations end up with multiple data sources for many good reasons, including independent departmental purchasing decisions, mergers and acquisitions, and SaaS services with native cloud storage.

ETL Adds Cost, Complexity, and Delays

To meet operational reporting requirements and ensure that accurate information is accessible to business users and analysts, data engineers frequently find themselves developing workflows that extract, transform, and load data (ETL). ETL workflows automate

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

data extraction from operational systems and land data in intermediate staging areas. From there, data is typically cleansed and transformed before being loaded into a data warehouse.

Data engineers have historically spent considerable time and effort carefully crafting data models that can answer anticipated business questions with reasonable performance. Tables are optimized to handle the queries typical of operational reporting applications and BI dashboards efficiently. ETL workflows are necessary to translate and reformat source data to match these target data models.

ETL workflows are also used for other purposes, including creating performance-optimized data views and extracts or building OLAP cubes for offline analysis. They may also periodically reaggregate or purge data in the data warehouse and move it to archival storage. A typical site may have hundreds of such long-running processes, often running during overnight batch windows.

Because ETL workflows run only periodically, the information in the data warehouse that feeds reports and dashboards is frequently out of date. Not only are these workflows expensive in terms of time and resources, but they also pose a substantial maintenance burden. Workflows are frequently brittle, and whenever a data source or schema changes, they need to be updated and validated.

Because of this maintenance burden and the associated data engineering backlog, analysts and business users often need to wait weeks for new or changed datasets. This represents a high cost in terms of lost productivity to the business.

Today's Environments Are Increasingly Hybrid

Data warehouses can deliver excellent query performance, but they are expensive places to store data. With data volumes growing, most organizations have sought more cost-effective ways to store and analyze vast amounts of collected data. Today, most organizations deploy data lakes alongside their enterprise data warehouse. Key business and financial reporting applications still rely mainly on the data warehouse. However, other datasets of interest to analysts and data scientists increasingly reside in the data lake.

This separation of data is a headache for data teams and analysts alike. Data is frequently moved and replicated across platforms by complex workflows. This results in the same data being stored in multiple places, often in different formats, complicating data governance. Also, organizations often use different query and analysis tools to access data in each platform, making it difficult to get a consolidated view of data and a single source of truth. A sample hybrid data lake/data warehouse environment illustrated in Figure 2 represents what many organizations are dealing with.

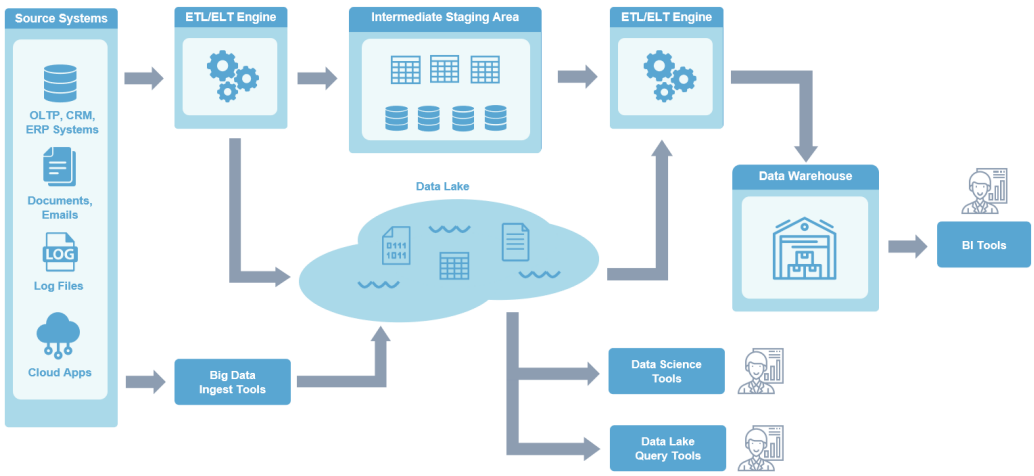

Figure 2 – Hybrid data lake/data warehouse with complex ETL/ELT workflows

Fortunately, a modern unified platform for data and analytics sidesteps much of the complexity associated with existing and traditional data pipelines.

Incorta – A Better Solution for Data Analytics

Incorta is a unified data analytics platform (UDAP) designed specifically to address these challenges. It does so in two ways:

First, Incorta is an all-in-one solution bringing together data acquisition, data processing, data curation, semantics, and data analysis, all accessible via a single web interface. Incorta preceded the current full-stack analytics platform trend by several years. Full-stack analytics platforms are a new breed of data analysis solutions that help organizations quickly get more value from their data. They deliver improved productivity and agility while significantly reducing the cost and complexity of the analytics environment.

Second, unlike a traditional database, data warehouse, or business intelligence application, there is no need for expensive and time-consuming data modeling and ETL operations to transform and reshape data. Incorta leverages its extreme performance advantage to enable business analysts, decision-makers, and data scientists to work together and efficiently analyze large and complex datasets in real time.

Incorta leverages two critical technology advances in database design to deliver dramatic performance improvements: columnar storage and an in-memory query engine. By organizing data in a columnar fashion, data can be compressed more efficiently, resulting in fewer disk seeks and more efficient CPU and memory use. Since most analytic queries deal with only a subset of columns, columnar storage can dramatically reduce the total volume of data read. Performance is further enhanced by caching data in memory, reducing I/O further and resulting in a faster, more agile, and flexible data warehouse environment.

Incorta also leverages a revolutionary proprietary technology called Direct Data Mapping™. Direct Data Mapping gathers information about structured data sources to deliver exceptional query performance, even with data sources having billions of rows, hundreds of joins, and thousands of tables.

This means that business users can easily access customizable, shareable data views and dashboards containing aggregated key performance indicators (KPIs) based on the latest data. They can also drill down to transaction-level data to answer business-level questions and gain insights into factors impacting the business.

Key Technologies

Before delving into Incorta's logical architecture, it is worth describing some of the core technologies behind Incorta. To help deliver the benefits described above, Incorta brings innovative new approaches to the challenge of building a unified data analytics platform. It also leverages the latest, state-of-the-art technologies that have arisen in the open source community.

Key Incorta Innovations

Some of the key technologies pioneered by Incorta are described below:

Direct Data Mapping

Direct Data Mapping (DDM) is Incorta's "secret sauce," lying at the heart of its exceptional query performance. With DDM, analysts can run sub-second queries joining multiple data sources without impacting performance. An enriched metadata map combined with smart query routing eliminates the need to pre-aggregate, reshape, or transform data. Analysts can collaborate on sophisticated analysis and generate visualizations within Incorta or access data residing in Incorta through a standard SQL interface.

Incorta Connector Architecture

Another Incorta innovation is its extensible connector architecture. Incorta can connect to virtually any database, enterprise application, data stream, or file format with 240+ connectors provided by Incorta and Incorta partners. A built-in Data Manager manages connections to multiple external data sources and allows data to be updated at user-definable intervals. Incorta's Data Loading Service implements various features, including parallel data loading, full or incremental loads, and support for data agents, allowing data to be gathered from behind corporate firewalls.

Memory-Optimized Analytics Engine

Incorta incorporates its own native in-memory Analytics Service that operates on data in compressed form. This innovative architecture delivers lightning-fast query performance for even complex queries involving multiple tables typical of BI workloads. The Analytics Service can scale horizontally across multiple on-premises or cloud-based compute nodes depending on business requirements, delivering near-unlimited scalability and performance. Analysts can select which Incorta tables are persisted in memory, providing fine-grained control over memory usage and query performance.

Business Schema — Self-Service Semantic Layer

Incorta features a built-in self-service semantic layer that allows analysts to create business-friendly data views. A business schema in Incorta represents a logical view of

underlying physical schemas tailored for specific reporting or analysis requirements. This capability improves security and dramatically simplifies providing data access to business users without the need for time-consuming data engineering activities and ETL processing. Users can easily create and share business-level views, subject to data governance controls. New columns can be added to business views using a flexible formula builder.

Incorta Blueprints — Pre-Built Analytic Content

Increasingly, organizations use commercial on-prem or cloud-based offerings for applications ranging from Enterprise Resource Planning (ERP) to Customer Relationship Management (CRM) to Sales Force Automation (SFA). Examples include Oracle e-Business Suite, Netsuite, SAP, JD Edwards, and Salesforce.com. Often, enterprises "reinvent the wheel," using third-party BI tools to access data in these commercial packages and ETL pipelines to create business-friendly views. Incorta Blueprints solve these problems, providing pre-built schemas and dashboards for accessing, organizing, and presenting data from popular business solutions based on best practices. Blueprints are covered in more detail later in this guide.

Advanced Analytics & Custom Visualizations

Incorta consolidates data management, analytics, and visualization into a single integrated platform, avoiding integrating multiple disparate systems. Analysts can easily create and securely share dashboards, reports, and insights with other users with minimal training. End users can choose to use the analytic and visualization capabilities native to Incorta, or they can access the same performance-optimized physical and business-oriented data views from popular BI tools such as Tableau, MicroStrategy, and Looker.

With dozens of customizable charts and multi-dimensional visualizations, analysts can create rich dashboards using an extensive library of visualizations. Everything from spider charts to bubble plots to stacked column and line charts to maps for presenting results by geography can be used to share meaningful business insights.

The Incorta visual framework is also extensible. The Incorta Component SDK allows developers to build unique visual and non-visual components, and share them with other members of the Incorta community through a marketplace. The Component SDK is based on the open source React library, the leading framework for building JavaScript web interfaces. This provides developers with limitless opportunities to create compelling, informative visualizations and custom applications.

Open Source Components

Incorta leverages leading open source technologies extensively in its design. By leveraging familiar, trusted, high-quality components, Incorta avoids reinventing the wheel to focus on higher-value innovation. The use of open source technologies also helps facilitate interoperability with third-party analytic tools and ensures flexible deployment options across clouds. Open source innovations leveraged by Inocorta include the following:

Scalable Cloud Data Lakes

While early data lakes were often built on Hadoop, today, they are more commonly deployed on scalable cloud object stores which are scalable, cost-effective, and can store any data type. Incorta can access data lakes implemented on POSIX file systems such as NFS or Amazon EFS, and supports Amazon S3, Microsoft ADLS Gen2, and Apache HDFS via native connectors.

Open File Formats

In the early days of Hadoop, organizations would frequently store large files such as logs or data extracts as text format files (CSV, TSV, JSON, etc.). Data teams quickly realized that this was inefficient. Text formats are highly compressible, often up to 90% or more, and data is more efficiently stored in compressed form. This is not a small consideration since a 90% reduction in file size leads to the same savings in storage costs. Also, without indexes, every query of a text dataset required that the entire file be read.

Various open file formats emerged to address these needs, including RCFile, ORC, Apache Avro, and Apache Parquet. While details vary, these file formats generally support data compression and deliver much improved performance. Parquet is a particularly efficient file format, purpose-built by Twitter and Cloudera to support analytic applications. Parquet stores binary encoded data in a column-oriented format and is widely used with platforms such as Spark, Impala, and Drill.[3] Incorta enables users to ingest data from all of these open files formats. Apache Parquet format files are used internally within Incorta because of their superior performance in supporting analytic queries.

Apache Spark

Apache Spark is an open source in-memory analytics engine for large-scale data processing. Spark was initially developed at the University of California at Berkeley's AMPLab to address limitations in Apache Hadoop MapReduce. A key architectural feature of Spark is its resilient distributed datasets (RDD) that redundantly store datasets in memory. Data residing in memory on a Spark cluster is physically persisted on disk in Hadoop file formats.

Spark's in-memory engine performs analytic operations in the order of 100x faster than MapReduce. It does this while maintaining compatibility with popular data stores such as HDFS, Apache HBase, Apache Hive, and Cassandra. Spark data stores are accessible using Spark SQL or standard ODBC/JDBC libraries.

Spark has emerged as a preferred platform among data scientists because it supports Python and R, languages widely used by data scientists, in addition to its native Scala. Spark is also popular because of its native machine learning library (MLlib). Spark provides

[3] Apache Impala and Apache Drill are SQL query tools designed for Hadoop to enable direct queries against data lake storage. Impala is shipped with Cloudera and MapR.

extensive built-in ML algorithms and workflows, making it convenient for data scientists to train ML models based on datasets accessible through Spark's DataFrame API.

Incorta bundles a full-featured Spark Engine managed within Incorta. Customers can easily create materialized views (MVs) that leverage Spark to efficiently query and enrich data residing in Incorta's shared storage.[4] Incorta also enables standard BI tools such as Tableau and Power BI to query Incorta's Parquet format data store directly via Spark through a PostgreSQL compatible SQL interface.

Interactive Notebooks

Increasingly, data scientists prefer to access code, data, and visualizations using collaborative notebooks. Notebooks provide an interactive environment where users can rapidly write and execute code, visualize results, and share code and insights with colleagues.

Incorta embeds a full-featured notebook interface based on open source Apache Zeppelin. Users can work within the notebook environment to explore, manipulate, and transform data and create materialized views. Once created, materialized views in Incorta are accessible to Incorta's analytic facilities and third-party BI tools. The Incorta notebook interface supports several languages and query environments, including PySpark (for Python users), R, Scala, Spark SQL, and PostgreSQL.

[4] Materialized views are a generic term referring to database objects that contain the results of a query. Materialized views are often used to cache precomputed results to optimize query performance in relational databases. In Incorta, materialized views are understood to refer to persisted derived views generated by Spark jobs.

Logical Architecture

Figure 3 provides a simplified logical diagram illustrating the major components in Incorta and how they interact with one another. Some Incorta services can optionally scale horizontally across physical nodes. For example, the Analytics and Loader Services can run across multiple hosts, either to provide high availability or additional capacity and scalability.

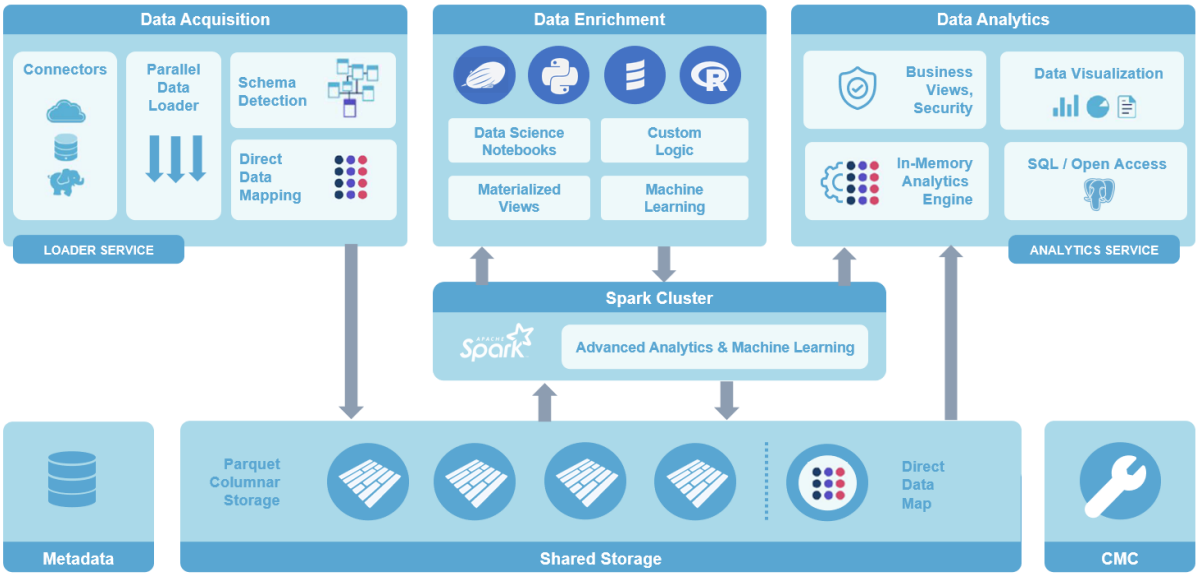

Figure 3 – A logical view of the Incorta unified analytics platforms

The following major components pictured in Figure 3 are included in Incorta:

- Loader Service (for data acquisition)
- Analytic Service (for data analysis)
- Data enrichment facilities
- Apache Spark cluster (optional)
- Shared storage (a data lake, or 'staging area')
- Metadata database
- Cluster Management Console (CMC)

In this section, we describe each of these components in more depth.

The Loader Service

The Loader Service in Incorta is responsible for extracting data from external data sources via connectors and persisting the data specified by physical schemas in shared storage[5]. It also maintains Direct Data Mapping (DDM) files and handles the enrichment and transformation of data by delegating processing to an optional Spark Cluster. The Loader Service is also responsible for loading data into memory as needed from shared storage.

The Loader Service is implemented as a Java application running on a Tomcat application server and Java virtual machine (JVM).

When loading data, Incorta leverages connectors to extract data from source systems. A Data Manager in Incorta (accessible from the Data tab in the web UI) allows tenants with appropriate permissions to manage external data sources and destinations. Each external data source uses a connector that specifies how the Loader Service ingests data. Data sources in Incorta can include databases, file systems, data lakes, streaming data sources, and even enterprise application platforms such as Salesforce, Splunk, or SAP ERP. Users can also create custom connectors using the Incorta Connector SDK.[6]

The Loader Service loads the data into memory (on-heap or off-heap) using connectors to build the columnar compressed Parquet and Direct Data Mapping files that the platform requires.[7] When the Loader Service loads new data and updates files in the shared storage area, the Analytics Service is alerted to which columns have changed, and it pulls changes into the in-memory analytics engine with no impact to users or applications.

The Loader Service does more than load data. It maintains access credentials, connects to each data source, and creates an inferred schema by gathering metadata, crawling the table structure, and recognizing data types for each column. The Loader service implements features that vary by connector. A partial list of these features are listed below:

- *Data chunking* – enabling data loaders to operate in parallel to improve data ingest performance by chunking data by size, date, or timestamp ranges.
- *Data agent support* – to load data via a proxy, enabling secure access to data on remote clouds or on-premises data sources behind firewalls.

[5] Sometimes referred to in the documentation and product as the 'staging area' or just 'staging'.

[6] Information about the Incorta Connector Software Development Kit is available at https://docs.incorta.com/cloud/references-connector-sdk/

[7] In Incorta, memory allocated to the Loader Service is configurable. In Java systems, on-heap memory refers to objects within the JVM memory subject to garbage collection. Off-heap objects in Incorta are stored in Ehcache, a caching solution widely used with Java-based services.

- *On-demand and schedule-based loads* – users can trigger data loads manually or use a built-in scheduler to load data into physical schemas.
- *Incremental loads* – rather than performing a full database extract, connectors can employ different incremental loading strategies, periodically fetching only data that has changed since the last successful extract or retrieving data based on columns containing numeric or timestamp values.

As Incorta executes an incremental load, it attempts to perform an "UPSERT" operation. The loader looks for a matching key for the entity being loaded, and updates the entity if it exists. If the key is not found, the loader performs an insert operation. This approach helps ensure data integrity and consistency between loads. Automated UPSERT functionality simplifies and streamlines the process of updating data in Incorta. It avoids the need for conditional logic that checks whether matching records exist before determining whether to update or create a record.

How Data Is Stored in Incorta

Before data can be loaded via a connector, a physical schema must exist to describe the ingested data. Incorta's Schema Wizard helps automate the process of building physical schemas. A physical schema can hold data from multiple sources and data from a source may be stored in multiple schemas. Suppose data is ingested from an external Oracle database. In that case, the Schema Wizard will create a physical schema within Incorta reflecting details such as table names, columns, primary keys, and relationships between tables reflecting metadata stored in Oracle.

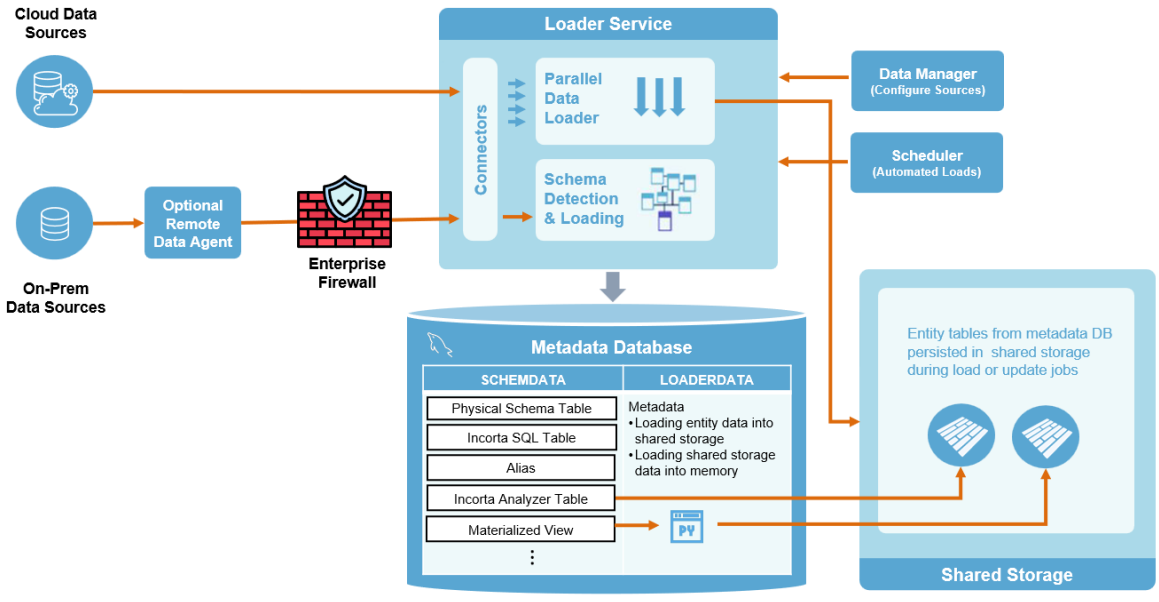

Figure 4 – Ingesting and storing data in Incorta

Figure 4 illustrates how data is ingested and stored in Incorta and the various components involved. Once a schema is defined in the metadata database, it can be used by the Loader Service and the Analytics Service to load and analyze data, respectively. Incorta users can use other tools in Incorta to review, refine, and annotate the physical schema, including a Schema Designer and Shema Diagram Viewer also built into the Incorta platform.

Despite its name, a physical schema in Incorta is an abstraction where each item in the schema comprises multiple components. Physical schemas are described using metadata stored in a relational database dedicated to metadata storage. The metadata defines the data imported into Incorta. This includes various entity objects, relationships between objects, and paths to the Parquet format files on shared storage that physically store the ingested data.

The metadata database contains details that govern how the Loader Service extracts, loads, and enriches data. It also controls how the Analytics Service loads data from the Parquet files into Incorta's in-memory analytics engine. Suppose a materialized view is created based on a particular physical schema (using Spark). In that case, a materialized view object will be added to the metadata database. The materialized view definition will reference the Python, R, Scala, or SQL code that creates the materialized view when executed on Incorta's internal Spark cluster.

Direct Data Mapping

In addition to storing metadata in a database and persisting the actual data in Parquet format files in shared storage, Incorta supplements this information with Direct Data Mapping (DDM) files also stored in Incorta's shared storage tier.

Incorta's DDM technology analyzes the source data schemas when data is ingested via the Loader Service. It determines which columns will be treated 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. This avoids the need for the complex ETL workflows described earlier to transform data into query-friendly formats. Incorta's unique Direct Data Mapping is a key reason that Incorta is both faster and more efficient than traditional data warehouse environments.

Incorta's Storage/Smart Data Lake Tier

Incorta does not perform analytic operations directly on external data. Instead, it first ingests data into its optimized shared storage tier, as illustrated in Figure 5. Incorta stores metadata in an internal relational database described above, and stores ingested data in Parquet format files accessible to the Loader Service, the Analytics Service, and the Spark cluster. Incorta also stores DDM files in the shared storage tier, providing additional metadata to speed operations.

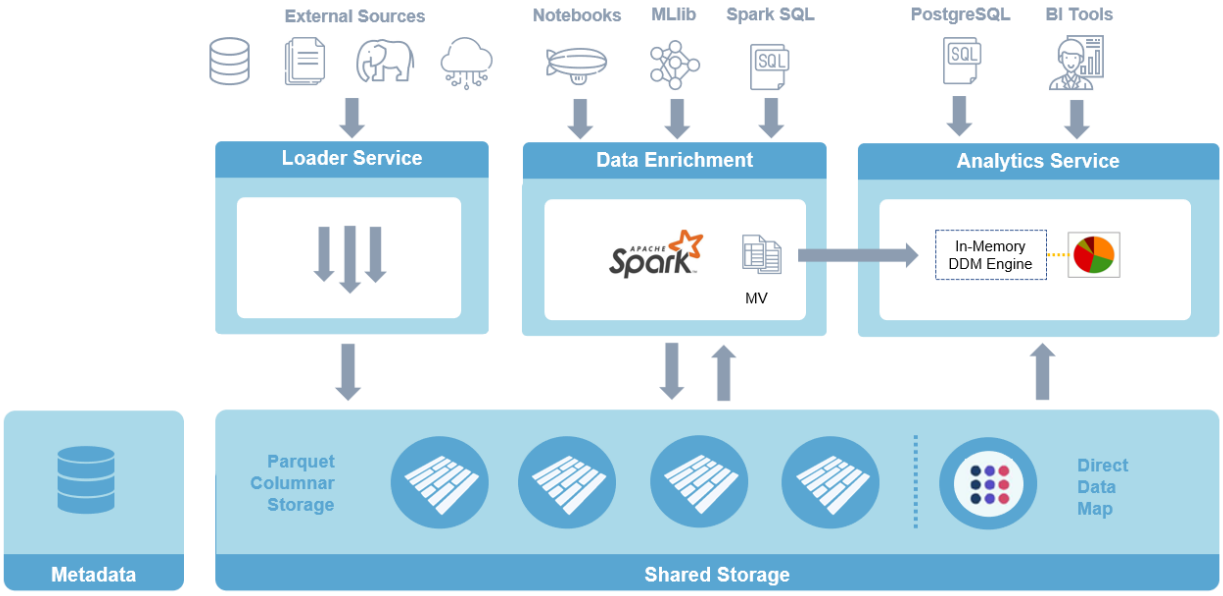

Figure 5 – Simplified view of Incorta's shared storage layer

Analysts have multiple options for accessing and analyzing data that resides in Incorta storage:

- Analysts can use Incorta's built-in analytic tools and dashboards.
- Data scientists familiar with Spark can analyze Parquet format data using familiar Spark tools and libraries.
- Business analysts can access various business and physical schemas in Incorta from third-party BI tools via a PostgreSQL compatible interface.
- Analysts can choose to bypass Incorta's in-memory DDM engine and access Parquet files directly using a separate PostgreSQL compatible interface listening on a different port.

The Incorta shared storage environment is implemented using a high-performance shared file system. Customers can elect to use standard NFS v4 or cloud file systems such as

Amazon EFS or Azure File Storage. Using a standard shared file system helps ensure that Incorta can be easily implemented across multiple clouds and on-premises environments.

Data Enrichment in Incorta

While most analysts will prefer to use Incorta's intuitive Analyzer or access shared dashboards in Incorta's content manager, other users have specialized needs. Apache Spark is a favorite platform for data scientists. Spark provides exceptional performance, easy-to-use APIs for operating on large datasets, and includes over 100 built-in operators for transforming data. Spark excels at iterative computation and includes libraries for statistical analysis, graph computations, machine learning, and SQL-based access.

Apache Spark performs most operations in memory and allows operations to execute on Hadoop data file formats, including Parquet. Spark also includes data connectors to multiple data sources. However, Incorta's built-in connectors are more capable and will be preferred by most users. The Spark-based facilities packaged in Incorta's data enrichment layer are illustrated in Figure 6.

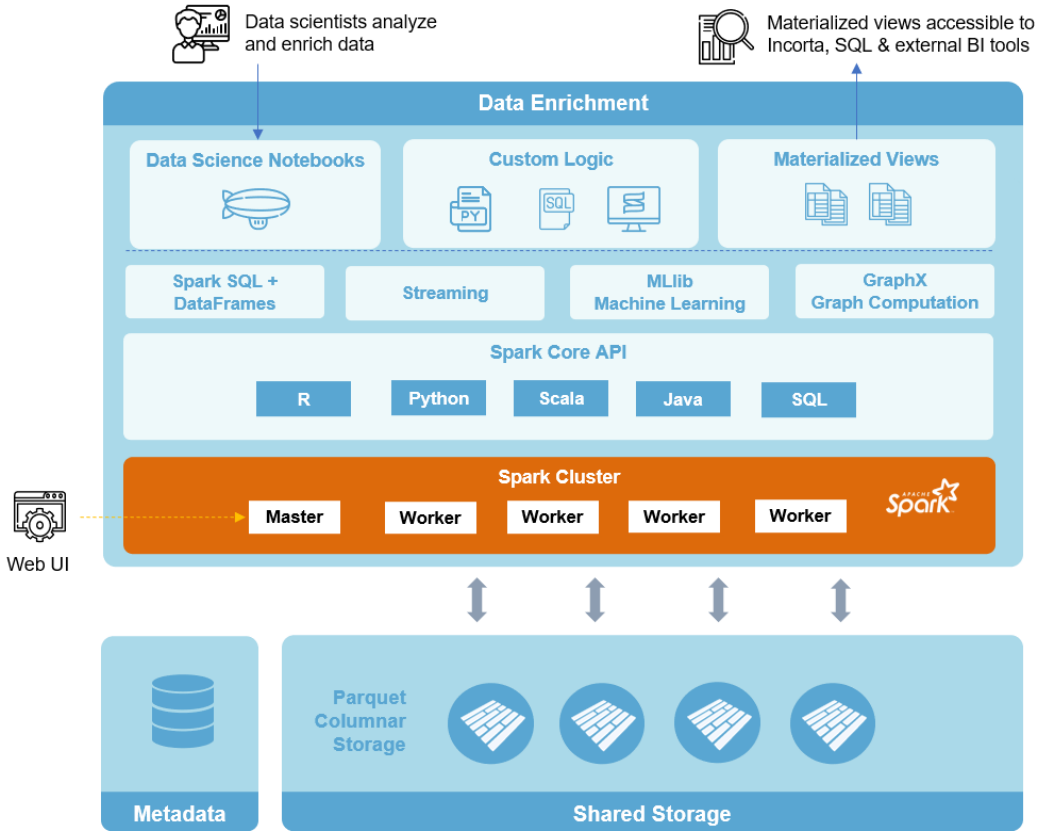

Figure 6 – Data enrichment in Incorta

Spark operations are accessed via Spark SQL or a data scientist's preferred programming language. Users can use language-specific libraries, including PySpark (Python module for Spark), SparkR (R frontend for Apache Spark), Java classes for Spark, and a Scala API.

Spark libraries translate high-level data operations into Spark jobs which are in turn divided into collections of tasks. These tasks execute in parallel across worker nodes on the underlying Spark cluster. Incorta exposes the Spark web UI enabling Incorta users to monitor the execution of Spark jobs and tasks. Monitoring Spark execution performance can help determine whether enough resources are assigned to the Spark cluster.

How Spark Is Used in Incorta

In Incorta, the use of Spark is optional. Although Incorta makes all of Spark's capabilities available, Incorta customers generally use Spark for two purposes:

1. Creating and executing Incorta materialized views to enrich data
2. Increasing the capabilities of the Incorta SQL interface
3. Executing parallel data extraction tasks

Data enrichment refers to the process of enhancing collected data with relevant data obtained from other sources. For example, suppose customer data is spread across multiple data sources. In this case, an enriched view of data may involve collecting data from each source, computing new values as appropriate, and adding them to a consolidated dataset. Data can be enriched in many ways to support different business requirements. Two examples are provided below:

- A lead management system may store a database table containing sales inquiries with details such as a prospect's name, phone number, and email address. By enriching this data with source IP addresses and device fingerprint information gathered from weblogs, analysts can score and prioritize inquiries based on browsing patterns and prior transaction history. For example, an organization may prioritize leads where a device user has made a prior purchase. Similarly, they may drop leads originating from devices tagged in a public RBL database as suspected of fraud.[8]
- A physical schema may store telemetry gathered from mobile devices storing latitude/longitude coordinates along with transactions. Analysts may wish to enrich this information by adding postal or zip code references and counties, cities, and states where operations occurred. This enables analysts to visually present geographic information on a dashboard and allow roll ups by county, city, and state.

The second use case for Spark in Incorta is increasing the capabilities of the Incorta SQL interface. The Incorta Analytics Service integrates with Spark via SparkSQL libraries included in the Spark framework. SQL queries may query Parquet files directly. Incorta passes queries that the in-memory DDM engine cannot handle efficiently to Spark for processing.

---

[8] Reputation Block Lists (RBLs) refer to public lists of domain names, URLs, and/or IP addresses that have been investigated and are suspected of fraud. These lists are fluid and are updated constantly.

The third use case is for parallel data extraction. This is to improve data ingest performance by accessing data according to size, date, or timestamp ranges.

Apache Spark Services in Incorta

Incorta users can use either a bundled Spark service or configure Incorta to use a preexisting Spark cluster. Most customers will prefer to use Incorta's built-in Spark services and avoid the complexities of cluster management and administering frameworks such as Apache YARN or Mesos common with Spark installations.

In addition to the use of Spark described above (enriching data with materialized views and accelerating SQL queries), users can also take advantage of native Spark facilities. Experienced analysts familiar with Spark can access Incorta data via Spark SQL. They can also use Spark for machine learning (MLlib), graph computation (GraphX), and other applications.

Incorta Notebook Interface

As explained earlier, notebooks enable data scientists to collaborate on exploring, manipulating, and transforming data. Incorta optionally exposes a full-featured notebook interface based on open source Apache Zeppelin. While analysts and data scientists can use notebooks for many purposes, they are commonly used in Incorta to create materialized views.

Incorta users can use the language-specific editor in the code section of a notebook to write code using PySpark or SQL and execute code interactively. Once a materialized view developed in a notebook is saved, they are treated as just another table by Incorta. Incorta can automatically refresh materialized views persisted on disk by re-running the Spark job against the latest ingested datasets.

Users that prefer to use external notebooks such as Jupyter or Zeppelin to access data within Incorta can leverage Incorta Data APIs for external notebooks. The Incorta Data API is a RESTful API accompanied by a Python library that allows Incorta data to be seamlessly read, queried, and saved from within a user's preferred notebook environment.

Analytics Service

The Analytic Service is a distributed service in Incorta that serves as the primary interface for users. Like the Loader Service, it is implemented as a Java-based service running on Apache Tomcat. The Analytics Service houses the in-memory Direct Data Mapping (DDM) engine. This engine answers all incoming data queries, whether through the Incorta UI or the Incorta SQL interface — both implemented by the Analytics Service.

The Analytics Service also implements a semantic layer providing secure access to business-level views. The Analytics Service also exposes a RESTful API through which developers can query data exposed via Incorta dashboards. A simplified diagram of the Analytics Service is provided in Figure 7.

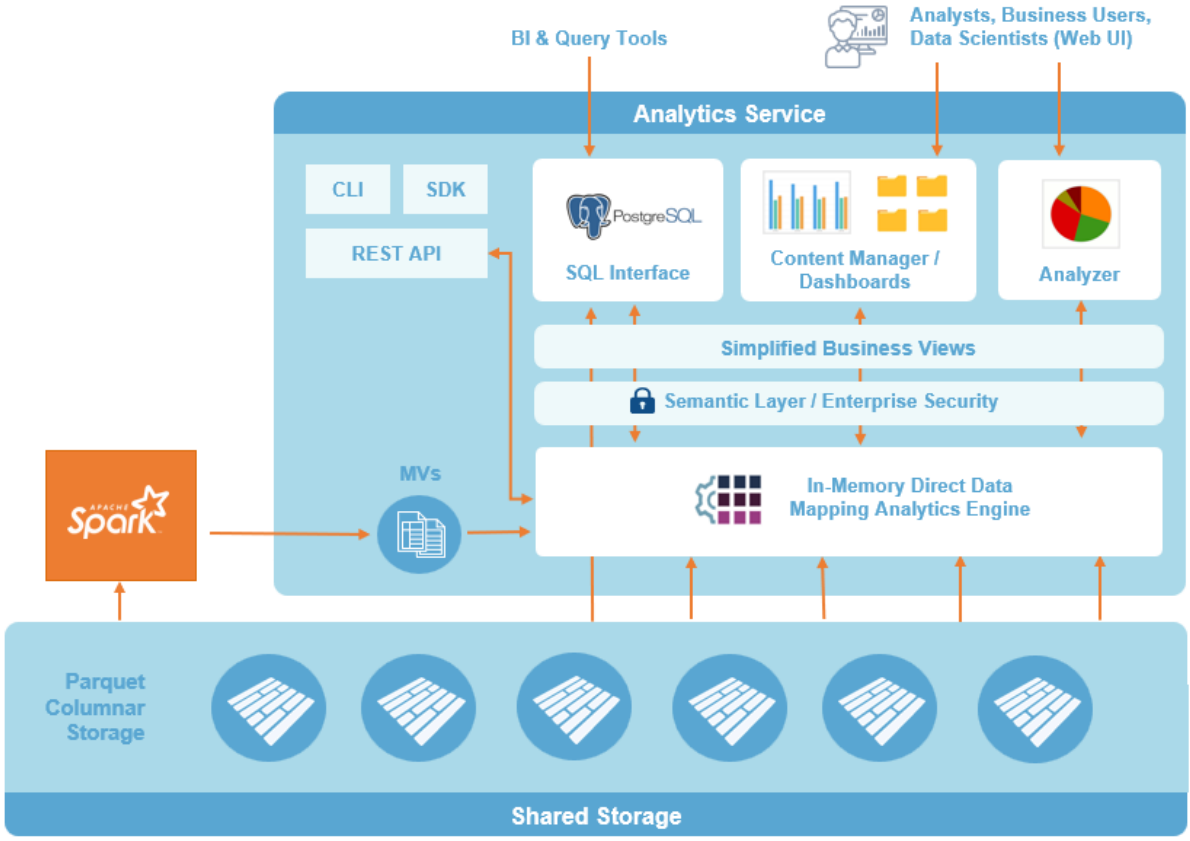

Figure 7 – The Incorta Analytics Service

The Incorta User Interface

Incorta provides analysts, data scientists, business users, and executives easy access to reports, dashboards, interactive visualizations, and analytics from their platform of choice.

Users can access Incorta's rich functionality from any device with a browser or take advantage of a full-featured mobile app available for iOS and Android smartphones and tablets. The mobile-native apps are built for power users and novices alike. They provide mobile shortcuts and other optimizations for small screens, along with an immersive augmented reality experience for iOS users.

Mobile interfaces are ideal for users such as line of business managers and field workers that need access to up-to-date information but may not always have access to a desktop computer. Regardless of the client device, Incorta provides users with a powerful and convenient way to access, view, and analyze business data from anywhere, anytime, with the ability to drill down into specific transaction details.

When users log into Incorta, they are presented with access to multiple Incorta facilities and tools. Users can navigate by selecting tabs across the top of the web interface or navigate the intuitive mobile app using a familiar mobile-optimized interface, as illustrated in Figure 8.

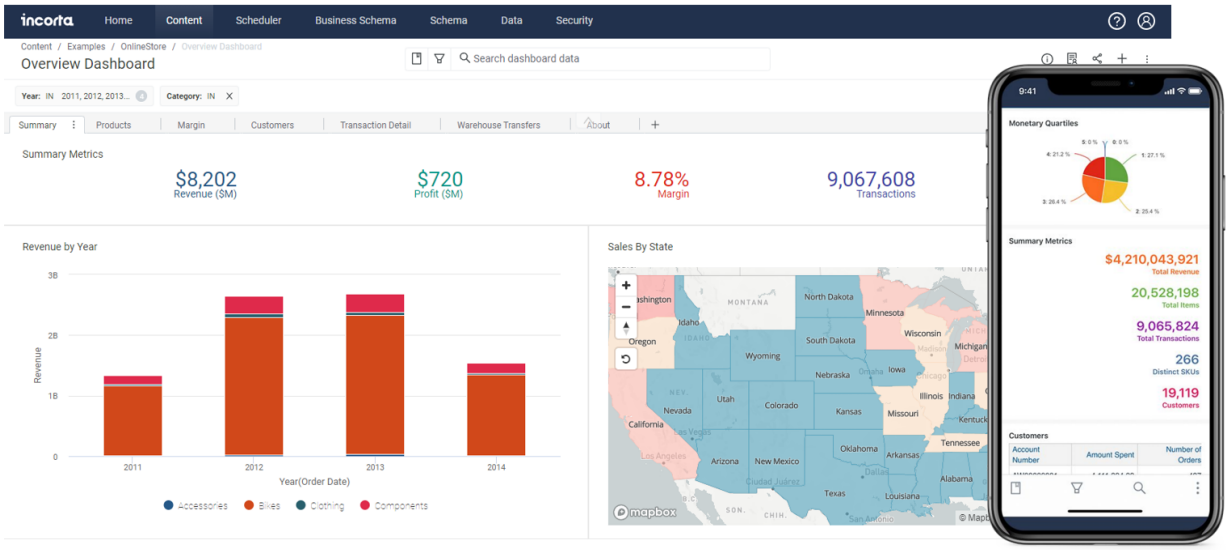

Figure 8 – Incorta provides a comprehensive UI that works with any device

Some key capabilities accessible through the web UI are described below.

Content Manager and Dashboards

Users can access pre-built dashboards through an integrated content manager. Dashboards can be placed in logical folders in the content manager and organized in hierarchies representing the business context. Through the content manager, users can selectively share both dashboards and folders.

Incorta dashboards are very powerful. Users can create customized dashboards that tell a story about the underlying data and easily share them with others while respecting security and data governance controls. Even analysts with minimal working knowledge of Incorta can become productive quickly.

Dashboards can support multiple tabs, rich text components, and users can easily drill down on data or set filters. Dashboards can also be made self-documenting so that non-specialists can easily understand them.

Dashboard authors can include custom components such as range sliders for interactivity or choose from dozens of pre-built visuals ranging from bubble maps to spider charts to maps with customized tooltips and popups. They provide practically limitless ways for analysts to present business data and share insights in new and interesting ways.

Figure 9 illustrates how Incorta Analyzer users can quickly create or edit a simple insight named “Revenue by Year” via the drag-and-drop interface. This insight breaks down annual

revenue by product category for an online store. Once created, the insight can be included in one or more dashboards helping users easily visualize and understand business metrics.

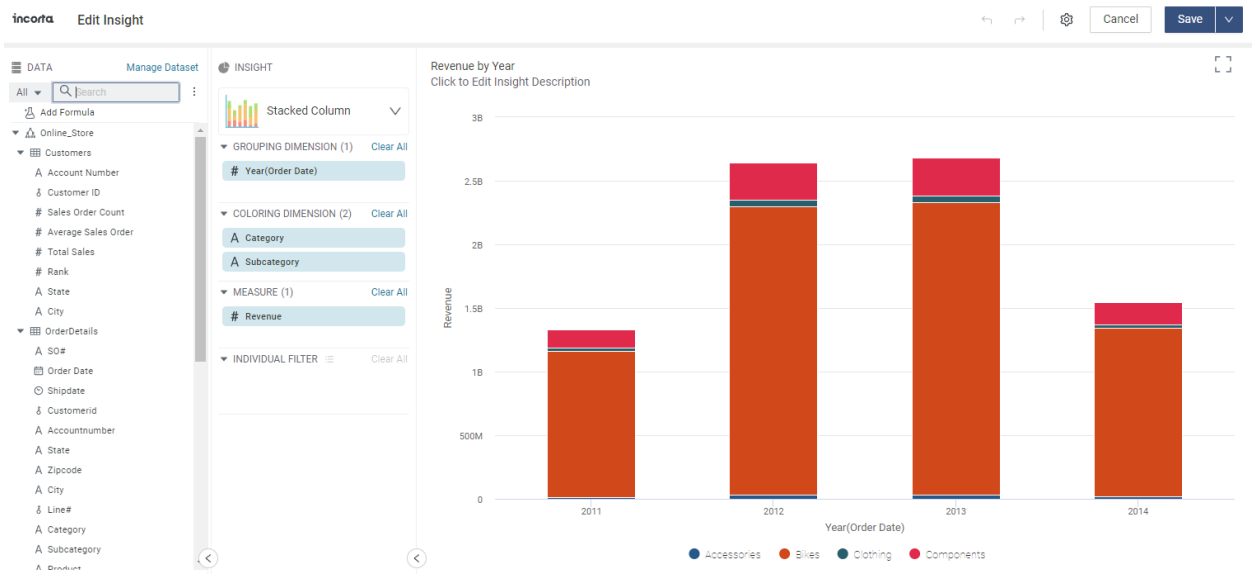

Figure 9 – Building insights using Incorta Analyzer

Like physical schemas, dashboards support versioning. Incorta users can quickly revert to prior views of a dashboard or work on modifications to existing dashboards without committing them to production. Users can preview, restore, export, or create copies of dashboards at any time, depending on their permissions.

For convenience, Incorta automatically recommends content specific to each user on Incorta's main page via a recommendation engine that leverages Incorta maintained metadata. This is a time saver that makes analysts more productive by avoiding the need for users to search for recently accessed content.

The Incorta Scheduler

Access to scheduling functionality is provided through the Analytics Service web interface. Using Incorta's scheduler, users can automate tasks such as loading or updating Incorta's internal data lake from source data. Load jobs can be defined to run automatically at configurable intervals.

The scheduler can also be configured to send regular dashboard updates or alerts triggered by changes to data in selected dashboard views. Automated email notifications can be sent based on configurable criteria in multiple rich content formats, including HTML, Excel, CSV, and PDF. This way, executives, analysts, and business users that never log into Incorta can still benefit from timely access to important information about the business.

The scheduler can also automatically trigger data updates to external cloud services. This is another powerful way to selectively share information with people inside or outside the organization that don't have access to Incorta dashboards. For example, the scheduler can automatically write CSV or Excel format files to a selected Google Drive or update a spreadsheet maintained in Google Sheets at periodic intervals.

Physical Schemas

The schema manager provides access to the various schemas maintained in Incorta. Users can create or import new schemas or use a schema wizard to create schemas based on external data sources available to Incorta.

Users can create new tables based on external data sources, joins that express relationships between tables, aliases, or tables derived from other tables in the physical schema. In Incorta, there are three types of derived tables:

- *Incorta Analyzer tables* – Special Incorta tables built from underlying physical or business schemas using the Analyzer tool. Incorta Analyzer tables are persisted to shared storage as DDM files rather than Apache Parquet files.
- *Incorta SQL tables* – SQL tables are table views derived from physical schema tables and created using an SQL select statement. SQL tables provide users with an easy way to create custom data views using familiar SQL syntax.
- *Materialized views* – As explained previously, users can write Spark code in notebooks using their preferred language (PySpark, SparkR, SQL, etc.) to generate customized or enriched materialized views. Materialized views are stored as Parquet format files in shared storage.

Other capabilities can be accessed through the schema manager as well. These include:

- Performance optimization features governing which tables are persisted in Incorta memory and which are stored as Parquet files in shared storage
- Version history, enabling users to create draft schema versions or revert to a prior version of a schema in the case of an error
- Interactive data exploration and ad hoc queries using the Analyzer tool
- Controlling the load order for various schema tables and materialized views
- Exploring data via an intuitive schema diagram viewer

Incorta supports data versioning for physical schemas. In case of a problematic update or user error during data loading, analysts or Incorta administrators can quickly revert to prior versions of a schema. Users can preview, restore, export, or create copies of prior physical

schemas at any time, depending on their permissions. This approach helps ensure data integrity and consistency between loads.

Administrators can also control the number of prior versions maintained for each entity and the backup frequency. Past versions of physical schemas can be exported to zip files. The ability to archive and restore past data schemas helps organizations comply with regulations that require data retention or calculation reproducibility, such as SOX and Basel.[9]

Business Schemas

Business schemas allow users to create business-friendly views of data from underlying tables without compromising security or requiring the space that physical tables would otherwise require.

Users can create business schemas using the Incorta Analyzer interface, formula expressions, or dragging columns from an underlying physical schema. Like Incorta dashboards, business schemas can be logically organized into file folders.

Figure 10 shows an example of creating a simplified business schema within Incorta. Rather than providing users with access to the dozens of tables that comprise the schema for an online store, a "SimplifiedRevenueView" exposes a smaller set of relevant columns for reporting or analysis. Like other objects in Incorta, business schemas can be shared selectively with other users.

Business schemas play an important role in helping organizations protect data, and comply with regulations such as GDPR and CCPA which govern the protection of personally identifiable information (PII). By sharing business-level data views, Incorta administrators can limit who has access to sensitive personal information.

[9] Sarbanes-Oxley (SOX) requires that organizations retain selected records for between 3 to 7 years and indefinitely for some datasets. Other regulations require that calculations be "reproducible" meaning that both data as well as analytic tools used to analyze data need to be retained.

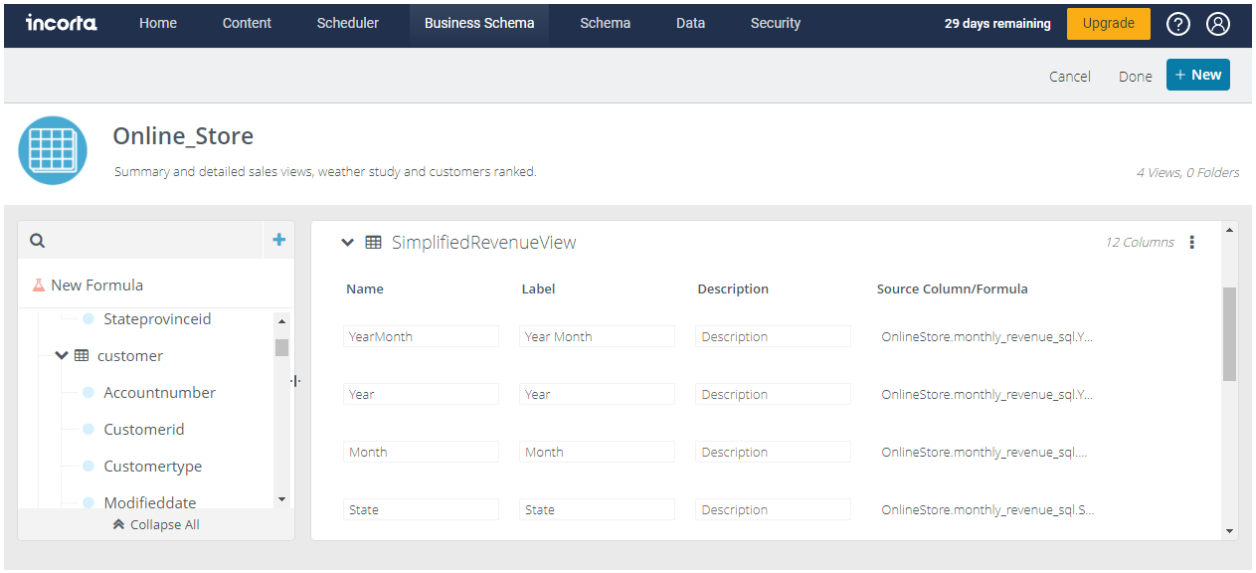

Figure 10 – Creating a simplified business schema in Incorta

Formula Columns, Formula Editor

Incorta supports the notion of formula columns in various types of Incorta tables, including physical schema tables, Incorta analyzer tables, business schema views, and within Incorta Analyzer. Formula columns are a powerful concept in Incorta. Formula columns are constructed using a Formula Builder built into the Incorta interface. Using the Formula Builder, users can derive or modify content in Incorta tables and custom views using rich expressions and custom logic.

The Formula Editor exposes a rich programming environment complete with code completion, syntax checking, and conditional logic. Users can easily assign column values based on complex logical expressions using this editor and its built-in domain-specific language. Expressions can reference built-in functions, various system variables, session variables, content from other Incorta tables, and more.

Data Sources and Destinations

Incorta users can manage data sources and destinations by selecting a Data tab across the top of the web interface. Data sources and destinations include:

- *External data sources* — accessible through data connectors provided by Incorta and Incorta partners.
- *Local data files* – Incorta provides each user with local server-side storage space accessible through the Incorta web UI. This space resides in a private directory on the shared file system. Users can manually upload and share folders or files in multiple formats, including Excel, Parquet, Optimized Row Columnar (ORC), and various text format files.

- *Data destinations* – Incorta can send Insights created through the Analyzer interface to directories or files in third-party cloud services listed as data destinations. Insight types including “listings tables,” “aggregated tables,” and “pivot tables” can be exported to services including Google Drive, OneDrive, or Google Sheets.

Cluster Management Console

The Cluster Management Console (CMC) is a web-based interface that administrators can use to manage an Incorta installation. The CMC is pictured in Figure 11. Through the CMC, administrators can manage infrastructure services such as the metadata database, ZooKeeper, Spark, and Incorta services, including Analytics Services, Loader Services, and notebooks. Incorta administrators can use the CMC to:

- Start, stop, or restart Incorta services
- View or update services or cluster configurations
- View the real-time state of services and clusters
- Create and manage cluster tenants and per-tenant configuration details
- Manage details related to SQL interfaces, the Spark integration, and data agents
- Retrieve detailed logs for each Incorta service for each cluster and tenant

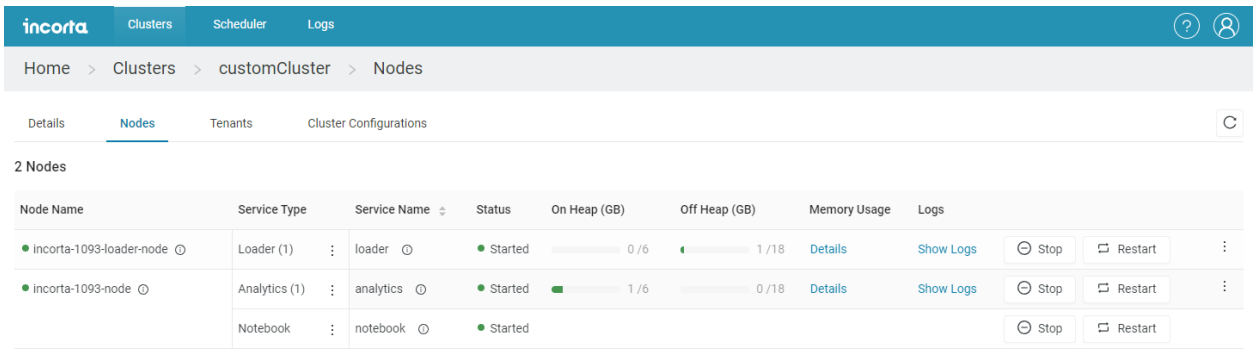

Figure 11 – The Incorta Cluster Management Console

The CMC also provides a scheduling service that backs up tenant configuration details periodically. The backup includes everything needed to restore an Incorta tenant, including dashboards, schema definitions, scheduled jobs, data sources, and users.

The CMC can also back up records of data lineage, consistency checks, and the current state of tables, schemas, business views, dashboards, and session variables based on a configurable schedule. These records can then be imported into the tenant for analysis, simplifying compliance and audit activities.

Incorta log files by service can be summarized into downloadable JSON or CSV format files for downstream auditing tracking all service events, including system errors, tenant competitions, service restarts, tenant schemas, and more.

The Incorta Security Model

Incorta employs a multi-tenant model. A single Incorta server or cluster can support multiple tenants, each with its own set of users, schemas, and dashboards. A “superuser” account administers each Incorta tenant. Tenants are separate, siloed entities that do not share data.

By default, each Incorta cluster has a default tenant. Users can access multi-tenancy features using the Cluster Management Console or a tenant management tool (tmt) to create and manage active tenants on an Incorta cluster. Multiple tenants can be configured on the same cluster to create environments that are logically isolated from one another with separate logins. For example, “dev,” “QA,” and “production.” Once configured, each tenant will have a distinct URL.

Security Roles

Incorta defines a set of preset roles for different types of users and administrators. User-level roles include User, Privileged User, Dashboard Analyzer, Individual Analyzer, and Analyzer User. There are also three different levels of administrative roles — Schema Manager, User Manager, and SuperRole. Each of these roles has different levels of privileges when administering Incorta entities.

Administering Users and Groups

Users are administered through the Security tab in the Incorta analytics interface. In order to add users, a user must be a member of a group with “Security” privileges. With appropriate permissions, users can create groups and assign security roles to groups. Users may be members of multiple groups, and groups may contain multiple users. Users can only assign privileges to others that they have themselves. For example, suppose a user is a member of a “Cloud Super Admin” group with full permissions on a cluster. In that case, the user can create another user and assign them to the same privileged group,

Assigning Roles to Groups

Permissions in Incorta are managed by assigning role definitions to groups and placing users in those groups. Incorta manages permissions using the following entities within Incorta:

- Catalogs
- Schemas
- Data destinations
- Data connections
- Security

For each of these entities, users may have “view” permission (shown in green), “share” permission (shown in orange) or “manage” permission (shown in purple). These permissions build on one another. For example, a member of a group that can manage catalogs can also share and view them.

A default set of permissions associated with each role in Incorta Cloud is shown in Figure 12.

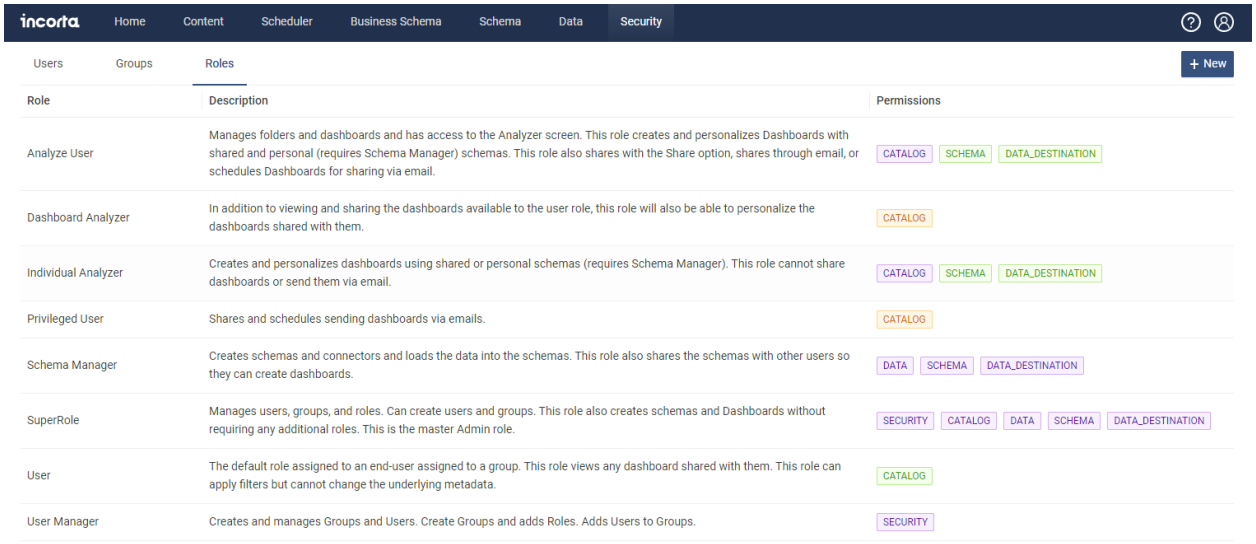

Figure 12 – Assigning Incorta roles to groups and users

Sharing Data in Incorta

When users create Incorta objects such as tables, business schemas, materialized views, or dashboards, they are private by default. Objects can be shared at the discretion of an Incorta user with various combinations of users and groups. Users can also control what permissions their colleagues have on objects that are shared. For example, objects may be shared as “view only” or “editable,” and users can control whether individuals can “re-share” objects. These fine-grained permissions combined with Incorta’s governance controls help ensure that access to data is appropriately controlled.

A clear advantage of Incorta is that security policies can be implemented in a single location for analysts enterprise-wide. Permissions set in Incorta apply to users of dashboards, Analyzer users running ad hoc queries, and users of third-party BI applications accessing Incorta objects through the PostgreSQL interface.

Row-Level Access Controls

In addition to providing security controls at the level of Incorta objects, Incorta can also enforce row-level access controls. For example, sales managers may need access to an “orders” table in an ERP system. However, the organization may only want managers to see orders with a “ship to” address corresponding to their sales territory.

To support this functionality, Incorta provides runtime security filters attached to various types of in-memory Incorta tables. Runtime security filters can control the rows visible to different users of the same table at runtime. The optional filter contains an expression that returns a Boolean value. If the filter evaluates to true, the row is returned. However, if the expression evaluates to false, the row is invisible to the user or downstream client application.

Runtime security filters can be simple or complex. For example, when evaluating whether a particular sales manager can access a row containing an order for “Western Region,” a two-line filter expression may check a separate “Territory Assignments” table to see if the active user (stored in a “$user” system session variable) is assigned to that territory. If they are, users have visibility to the row. Otherwise, the row is not returned.

This concept of runtime security filters is powerful because it avoids creating separate physical tables or data views for different analysis requirements. Also, role-based and row-based access controls are enforced whether the query is made from within Incorta or by third-party BI tools accessing Incorta through the external SQL interface.

Column-Level Access Controls & Data Masking

Data masking restricts access to data in sensitive columns. Consider the case where an organization has a central employees table. The table may be used for several different dashboards and reports. However, some of the information in the table is confidential. For example, personal contact information, job-level classifications, and salary and benefits-related information all need to be protected. Columns containing confidential information should be visible only to human resources or executive management group users.

Users can take the same approach as above and apply runtime security filters to mask sensitive columns. Conditional logic can be used to mask particular data columns depending on the user making the query. Users may see no data returned for masked columns or see a placeholder value such as "*****" depending on the organization's preference. Information may also be selectively obscured, as is frequently done with email addresses or credit card numbers.

Data Encryption

By default, Incorta encrypts all sensitive data, including user passwords, passwords to external data sources, and security tokens. Users can also select whether data stored in Parquet format on disk is encrypted by using controls exposed through the schema interface. Once data is in memory, security is enforced by the Analytics Service. Incorta encrypts data using the Advanced Encryption Standard (AES 128).

User Authentication

Manually managing user accounts within Incorta can be tedious, so Incorta supports external authentication services in addition to its native password authentication. Single sign-on

(SSO) options are configurable for each Incorta cluster on a per-tenant basis. Incorta is compatible with SSO providers that support the Security Assertion Markup Language 2.0 (SAML2) protocol, including:

- Okta – Okta Identity Platform
- Auth0 – Auth0 authentication service
- ADFS – Microsoft Active Directory Federation Services
- IBM CIS – IBM Cloud Internet Services authentication
- OneLogin – OneLogin cloud authentication

Incorta also supports the Lightweight Directory Access Protocol (LDAP) as a central source for authentication.

Compliance

Incorta provides robust security controls and is continually working to enhance security and ensure that data is protected. Incorta Cloud has established operational controls in accordance with SOC 2, a robust set of cloud security controls established by the American Institute of Certified Public Accountants (AICPA). A SOC 2 Type II report, produced by an independent auditor and updated annually, is available on request.

Incorta has also taken best practices measures to ensure compliance with the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Incorta provides a wide variety of security capabilities described above to help organizations comply with additional regulations and safely care for and handle personally identifiable information (PII).

Physical Architecture

Customers can elect to deploy all Incorta services on a single node or deploy Incorta in a highly available clustered environment. Most production environments will run a clustered environment where Incorta services are spread among multiple hosts, improving performance and availability. Regardless of how and where Incorta is deployed, Incorta runs the same set of software services.

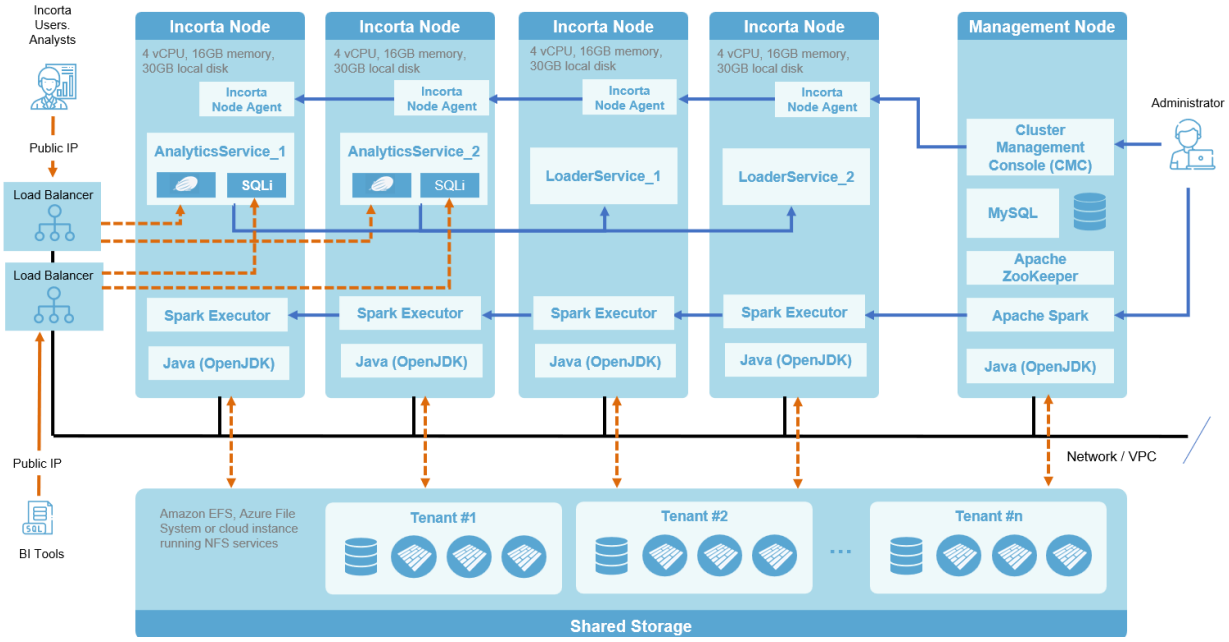

Figure 13 – Sample multi-host Incorta deployment

The physical architecture for Incorta is not set in stone. There are multiple potential deployment models ranging from single server deployments to the highly available architecture illustrated in Figure 13.

Some customers may be concerned about business continuance and deploy a separate Incorta environment for disaster recovery in a separate data center or cloud region.[10] Other customers may use an existing Spark cluster or a separate relational database to host the metadata database.

The sample deployment architecture illustrated in Figure 13 is described below:

- A management node runs key Incorta services, including:
    - *The Cluster Management Console* – For monitoring and managing other cluster nodes and ensuring the availability of Incorta services.

[10] See the Incorta Best Practices article on Disaster Recovery by Mateen Mohammed – https://community.incorta.com/t/q6hf97c/disaster-recovery

- *A relational database housing the metadata database* – For small implementations, a bundled Apache Derby database can be used. MySQL or Oracle databases are recommended for production installations. The metadata database may be architected for high availability in production, either using MySQL high availability features or a highly available MySQL compatible cloud service.

- *Apache Spark master and Spark web UI* – For monitoring and management of Spark jobs.

- *Apache ZooKeeper* – Used to coordinate messaging between components and elect new leader services in case services fail. ZooKeeper runs across all nodes in the Incorta cluster but is only pictured on the master node in Figure 13. ZooKeeper requires an odd number of hosts (to facilitate elections and keep quorum with an ensemble), so this is an important consideration when planning a custom Incorta deployment.

- A clustered deployment will typically have multiple Incorta nodes depending on load and availability requirements. Additional Incorta nodes support the following services:

  - *Node Agent* – A Node Agent is installed on each Incorta node. The Node Agent is present even in single server installations. The Node Agent is part of the Incorta HA package and monitors Incorta services and starts and stops nodes.

  - *Analytics Service* – Each Incorta node may support one or more instances of an Analytics Service. The Analytics Service is typically installed on a dedicated host or VM. The Analytics Service provides the main interface in Incorta for analytic users. The Analytics Service also implements a PostgreSQL compatible interface to access Incorta-resident data. Instances of Analytics Service are managed using the web-based CMC on the management host.

    When serving large user communities, it is helpful to use machine instances with lots of memory and multiple cores. The Analytics Service is both vertically and horizontally scalable. Requests may be spread across multiple analytics engines using a load balancer, as illustrated in Figure 13. Running multiple Analytics Services increases scalability and also ensures availability.

  - *Loader Service* – Like the Analytics Service, the Loader Service can run across one or more hosts in an Incorta cluster. The Loader Service will store metadata in the MySQL-based metastore and write Parquet format files to the shared storage area. Nodes running the Loader Service are typically provisioned with large amounts of memory. Keeping data in memory helps the Loader Service merge data from incremental loads more efficiently before writing data to the storage layer.

Although the Loader Service can be deployed on the same host as the Analytics Service, there are advantages to running it on a dedicated host (or hosts). Some environments may schedule large or frequent data updates. Separating the environments helps ensure that the analytics environment remains responsive. The Loader Service can execute compute and memory-intensive tasks such as rebuilding table joins or recalculating formula columns during updates, even when the datasets updated are relatively small.

- *Shared storage* – Shared storage may be implemented as a shared NFS file system or using a POSIX cloud file system such as Amazon EFS, Azure Files, or Google's GFS. All Incorta hosts will need to mount the same shared storage. The shared file system will store files associated with all tenants. However, Incorta users will only have access to files for the logical tenant they are logged into.
- *Load balancers* – Incorta can provide added availability and scalability using load balancers. Load balancers allow users to access a single external IP address and port as an endpoint while distributing load across multiple servers or cloud instances.

Deploying Incorta in the Cloud

Although Incorta can be deployed across multiple cloud environments, most customers will prefer to use Incorta Cloud. Incorta Cloud is a fully managed cloud service based on best-in-class data and analytics infrastructure that deploys in minutes without requiring customers to have specialized technical expertise or manage cloud infrastructure.

Incorta Cloud is the recommended way to deploy your Incorta environment. Incorta Cloud offers several advantages, including:

- Users can start small and quickly expand capacity as needed
- Incorta Cloud offers virtually unlimited storage capacity in the cloud
- Superior performance for data loading, data analytics, and data science processes
- Services are configured following best practices for superior reliability
- A managed service offers greater security since credentials for internal Incorta services are handled within Incorta Cloud and generally not exposed to users and administrators

With Incorta Cloud, users can customize their deployment to meet their needs directly through the Incorta Cloud web UI. Incorta Cloud looks after all of the details associated with managing cloud infrastructure including details such as credentials, machine instances, cloud storage, virtual private cloud configuration, load balancers, and more. Through the Incorta Cloud interface, administrators can easily tailor the Incorta environment to their needs performing activities such as:

- Creating, managing, and removing Incorta clusters
- Creating and managing users accounts, and establishing user roles associated with each Incorta cluster
- Managing the number of nodes and the amount of memory allocated to the Loader Service and Analytics Service within each cluster to meet the throughput and availability requirements
- Accessing various URL endpoints and customizing functionality on a per-cluster basis such as the SQL interface, Spark-based data enrichment facilities, and the notebook interface
- Creating, managing, and removing logical tenants within each Incorta cluster

Incorta Uses and Practices

So far, we have described Incorta's overall architecture and explained how it is deployed in various on-premises and cloud environments. In addition to the capabilities described above, Incorta provides other capabilities as well. Some of these are described below.

Incorta Blueprints

Blueprints provide pre-built schemas and dashboards for accessing, organizing, and presenting data from popular business solutions based on best practices. Blueprints include schemas for multiple functional areas in each application. These business schemas pull together key metrics, sample reports, visualizations, and sample dashboards. Blueprints do the "heavy lifting" so that only light customization is required for local customers after installation.

The results are dramatic — speeding time to value by reducing the work necessary to deliver baseline reports and KPIs. Incorta also makes it easy to combine data from multiple third-party business applications for a single view of data across the enterprise. For example, users can create queries from sources such as Oracle E-Business Suite or SAP. They can copy these queries into a table and easily create a UNION of both tables, eliminating duplicate columns and identifying data sources for each column and how to resolve conflicts.

Incorta Blueprints are available in multiple disciplines, including operations, finance, supply chain, sales, and other cross-function applications. Some available Incorta business blueprints are listed below, and this list is growing all the time.

- Incorta Analytics for NetSuite — Business Intelligence Analytics
- Incorta Analytics for Salesforce
- Cash-to-Cash Cycle Analytics
- Accounts Payable Analytics
- Enterprise Asset Management Analytics (EAM)
- Inventory Analytics: Inventory Management Analytics Software Tools
- Incorta Analytics for Oracle E-Business Suite (EBS)
- Incorta Analytics for SAP
- Policy & Claims Analytics
- Accounts Receivable Analytics
- Fixed Asset Analytics
- General Ledger Analytics
- Order Fulfillment Analytics
- Procurement & Spend Analytics Dashboard
- Incorta Analytics for Oracle ERP Cloud – Cloud Analytics and Reporting Tools
- Incorta Analytics for Oracle JD Edwards (JDE)
- Supply Chain Planning and Analytics
- Bill of Materials Analytics
- Order to Cash (O2C)
- Procure to Pay (P2P)
- Tariffs Reporting and Analytics

Analyzing Incorta Metadata

In addition to providing blueprints for analyzing data from third-party applications, Incorta leverages this same powerful idea of pre-built blueprints to provide business-friendly views of Incorta's internal metadata. Incorta supplies two built-in "Incorta Utilities" blueprints with

extensive views and dashboards that make it easy to analyze and report on activities with Incorta. These blueprints are:

- Incorta Metadata
- Incorta Inspector

These blueprints provide capabilities that include auditing access to various Incorta objects, tracking data lineage, and helping ensure compliance with various data governance controls, organization policies, and regulatory requirements.

Once installed, dashboards from these blueprints appear in folders in a user's content area. These blueprints catalog all objects in the schema and track all user activities, including accessing and modifying dashboards, running queries, and accessing or creating other objects. Even query performance and failed queries are tracked so that Incorta administrators can identify and resolve bottlenecks or quickly determine why particular queries may have failed.

Machine Learning for Automated Insights

Because Incorta works directly with rich source data, it is an ideal platform for machine learning. Incorta provides machine learning capabilities via two distinct capabilities:

- *Built-in ML libraries accessible through Spark Notebooks* – Incorta provides a native Incorta ML library that acts as a wrapper over Spark ML and other machine learning libraries to simplify the process of creating, training, and running models.
- *Incorta Analyzer* – Incorta provides separate machine learning functionality built into Incorta Analyzer, making it easy for analysts or data scientists to create insights that use machine learning for prediction or anomaly detection.

The integrated notebook environment provides an ideal environment for developing and training machine learning models. Data scientists can work in a familiar notebook environment to develop and train models and then use that model to build materialized views that can be accessed within Incorta. As Incorta ingests new data, model predictions can be updated automatically.

Incorta's ML library is essentially a wrapper around Spark ML and other popular ML libraries that overrides specific methods to ease model development and training. For example, users can directly import Incorta objects into Spark DataFrames and save them to Incorta storage. Users supply an Incorta object name in their notebook code, and Incorta looks after the rest. Data scientists can use native methods in PySpark and other language environments or access Incorta enhanced implementations of those methods. For example, an Incorta supplied *incorta.printSchema()* library method provides a rich rendering of DataFrames imported from Incorta, as shown in Figure 14.

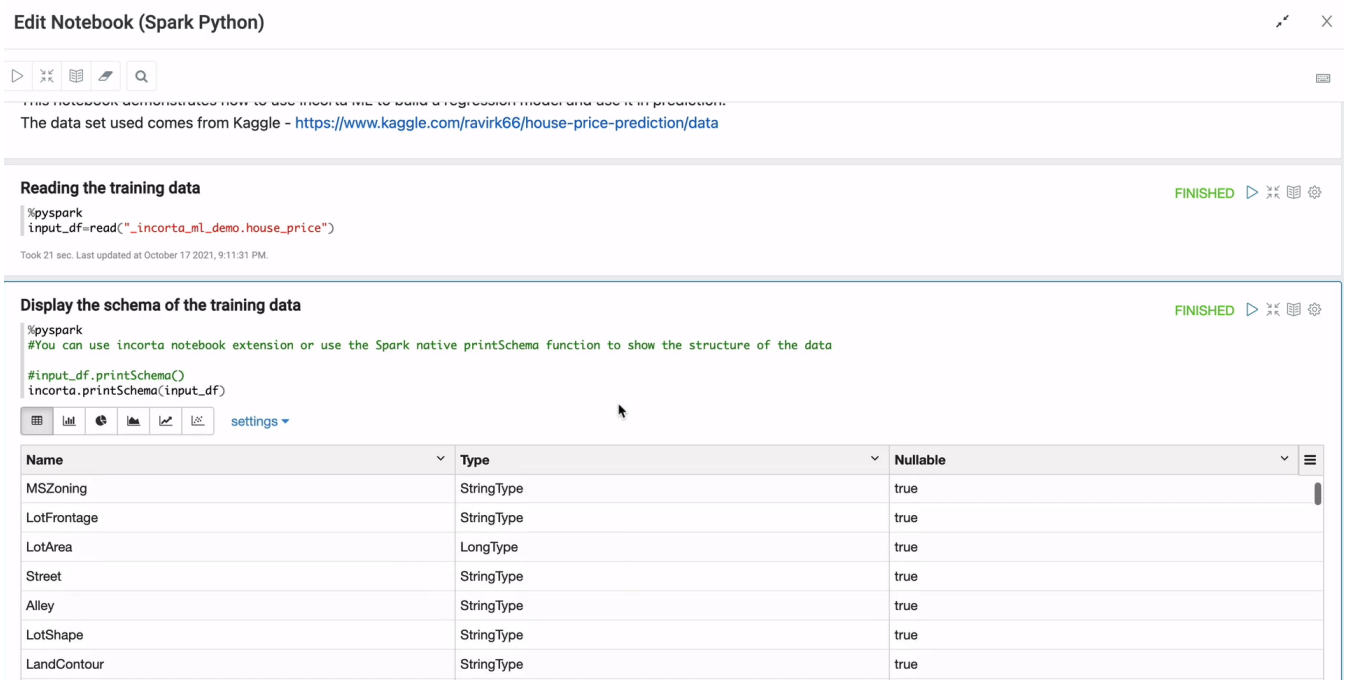

Figure 14 – Training ML models using the Incorta ML library

With Incorta ML, data scientists can easily prepare training data, perform feature selection, train and evaluate models, and run predictions. Prediction results can easily be saved back to Incorta schemas or saved as materialized views.

One-Click Machine Learning Using Incorta Analyzer

In addition to the rich ML capabilities accessible through Spark notebooks, non-technical users can use ML features built into Incorta Analyzer. Analysts can augment charts with data generated by pre-trained models for time-series forecasting and anomaly detection. These ML models drive insights that can then be incorporated into dashboards and shared appropriately across the enterprise.

For example, an organization may wish to predict inventory levels for various SKUs based on historical or seasonal purchase patterns. Others may wish to create visualizations that highlight unusual activities for further investigation that may indicate fraud or possible errors. Customers can select from multiple predictive models, including auto ARIMA, Facebook Prophet, and exponential smoothing.[11] Alerts may be triggered if embedded ML models predict unusual results.

[11] ARIMA refers to "auto regressive integrated moving average," a model in R and Python for predicting future data. Prophet is an open source forecasting model developed by Facebook. Exponential smoothing is another popular forecasting method for time-series data.

Data Governance Simplified

Data governance is a broad topic. It refers to the various processes, policies, standards, and metrics that organizations apply to ensure that data is managed effectively and securely. Organizations are typically concerned with managing data quality, understanding data lineage, implementing process changes, and creating a data catalog.

A key advantage of Incorta is that it enables data to be collected and analyzed in a single, flexible system. By storing data in a central catalog, data governance is simplified dramatically. Incorta also enables organizations to avoid time-consuming and costly ETL/ELT pipelines. Rather than relying on transformations, Incorta provides fast analytics on full-fidelity data.

Incorta's integrated Loader Service can replace the entire ETL infrastructure and perform full or incremental data updates regularly. Incorta's DDM technology avoids the need for complex data transformation and intermediate data copies that complicate governance. Fast and efficient data loading also helps ensure that decisions are made based on the latest available data.

Incorta delivers essential features that simplify data governance, including:

- *Table and dashboard versioning* – Incorta maintains multiple versions of dashboards and tables. New versions are automatically created when data is ingested. Users can name versions, roll back at any time, and dashboards and schemas can be easily exported and reimported later.
- *Business-level schemas* – Rather than copying data to meet user requirements, Incorta supports business schemas derived from data in underlying physical schemas. This helps avoid unnecessary data copies and ensures that analysts are working with the latest data.
- *Row-level access controls and data masking* – Incorta supports row-level access controls via flexible security filters and data masking to ensure that only authorized users have access to sensitive data.
- *A comprehensive metadata dictionary with pre-built dashboards for analysis* – Incorta provides a single data catalog that tracks all Incorta objects and system activities. Metadata in Incorta can be analyzed just like any other data source.
- *Data lineage tracking/impact analysis* – Using dashboards for metadata analysts, users can easily understand how objects are used in Incorta. For example, before changing a physical schema, or calculated field, users can determine which dashboards, reports, and derived views may be impacted by changes to particular columns.

- *Underlying data schemas can be modified without breaking dashboards* – The semantic layer in Incorta abstracts underlying data. Users can often make changes to data without affecting derived tables and dashboards. The impact analysis features described above help users understand when it is safe to make changes.
- *Extensive catalog search capabilities* – Every object in Incorta is searchable, including users, dashboards, physical schemas, joins, business schemas, and more. Metadata can be queried just like any other data type and placed on dashboards to facilitate search.
- *Comprehensive auditing* – All usage and development-related activities are tracked and can be reported on at any time with rich detail by user, dashboard, and query. Incorta writes a log of all user activities for each tenant to an audit.csv file that can be analyzed using pre-built dashboards or within Incorta Analyzer.

Working With Other Tools

Although Incorta provides a complete unified data analytics platform, it is portable, open, and extensible. Incorta supports both on-premises and cloud-based deployments, providing flexibility and ensuring that customers are not locked into a particular cloud service or infrastructure platform. Incorta also provides a variety of open interfaces so that existing customer investments are protected.

Support for Third-Party BI Tools

While most users will prefer to use Incorta's built-in Analytics Service, some customers may already be familiar with third-party BI tools and have libraries of pre-built dashboards. Incorta features built-in integrations for popular BI tools such as Tableau, Microsoft Power BI, and Microsoft Excel. Supported integrations can be enabled via the CMC. Once enabled, users are provided with guidance related to supported drivers, access endpoints, and clear instructions to enable each BI tool to access data in Incorta.

Users connecting to Incorta via third-party tools will enjoy superior query performance and access to more up-to-date data by accessing the same schemas available to Incorta users. Data governance is also simplified because access controls are enforced centrally within Incorta. With this functionality, organizations can easily migrate at their own pace. Users can continue to use familiar BI tools but bypass costly and complex ETL pipelines while taking advantage of rich datasets in Incorta that are sharable enterprise-wide.

Excel Add-In

Using an optional Excel add-in, users can easily extract data from Incorta into Excel tables and pivot tables. This allows users to save queries of offline use, and data can be refreshed from Incorta when users re-open their spreadsheet. Using this add-in, users comfortable with Excel's built-in charts can easily create meaningful visualizations and share insights based on data stored in Incorta.

Open SQL Interface

An open SQL interface (SQLi) serves as the basis for open data access in Incorta. The SQLi exposes a PostgreSQL-compatible interface to Incorta data that can be used to connect third-party query and BI tools. Any client that can connect to a PostgreSQL database can connect to Incorta. The SQLi can be toggled on or off by administrators using a "Connect External BI Tools'' option within Incorta.

The SQL interface listens on two different TCP/IP ports. One port enables users to run queries directly against Incorta's in-memory engine providing access to Incorta's derived views and business-level schemas. A second endpoint enables users to query underlying physical Parquet files using SQL. The integrated Spark engine in Incorta is leveraged to transparently accelerate these queries.

Scriptable CLI

For organizations that wish to automate operations in Incorta, a Python-based CLI is provided. The CLI can be used with scripts to automate management activities, including importing and exporting Incorta objects, managing roles, and managing user access. The CLI is accessible to users with shell access to a node where the Analytics Service is running. A series of sample scripts illustrating how to use the CLI is available in the Incorta *bin/* directory. By combining the Incorta CLI with similar CLIs provided by cloud services, users can automate a wide range of actions in Incorta.[12]

Incorta REST API

For developers, Incorta provides a public RESTful API. The REST API in Incorta is used to access data rather than administer the system. Developers can post JSON formatted messages to various Incorta API endpoints using their language of choice. The Incorta documentation provides examples showcasing the use of the REST API in Python, Java, JavaScript, and C#. REST endpoints are provided for token creation, token refresh, dashboard prompts (returning a list of a dashboard's dimension values to inform a query), and a dashboard query endpoint for making requests against specific dashboards.

Incorta Data API

An Incorta Data APIs enables users to access data stored in Incorta, run queries on the data, and save data back to Incorta from their preferred machine learning tools, including external notebooks such as Jupyter or Zeppelin. These RESTful APIs are accompanied by a Python library. Using the Incorta Data API users can keep data within Incorta, which provides a higher level of performance.

Open SDKs

In addition to the open interfaces described above, Incorta also offers open SDKs to extend the functionality of Incorta:

- The Incorta Connector SDK enables Incorta customers and partners to implement their own custom connectors. The SDK is implemented as a Java Archive (JAR) file included with the Incorta distribution.
- The Incorta Component SDK enables users to create custom visualizations using React, the leading JavaScript UI framework, and publish them to the Incorta marketplace, where they may be shared with other users. Once developed, custom visualizations can be accessed through Incorta Analyzer and incorporated into dashboards without the need for any custom coding.

[12] Most cloud providers offer similar CLIs to automate actions. Examples include the AWS CLI, the Azure CLI, and Google's gcloud CLI, all implemented in Python.

Summary

Incorta can provide a decisive advantage for organizations of all sizes. It provides a single, unified data analytics platform that can consolidate data management and analytics functions across the full range of enterprise data. 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.

Key advantages of Incorta include:

- Analysts and data scientists become productive immediately with rich, intuitive interfaces that make it easy to realize valuable insights from data
- A library of high-performance connectors enables Incorta to connect to any data source without the need for slow and costly ETL processing and associated infrastructure
- A complete suite of integrated tools support lightning-fast data analysis and visualization while enabling existing BI tools to query Incorta's high-performance data engine directly
- Incorta can be deployed in minutes as a fully managed service and is open and extensible, supporting deployments on-premises or on a customer's preferred public or private clouds
- A full-featured semantic layer provides self-serve data access with robust security and data governance controls

With Incorta, 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 https://cloud.incorta.com/signup.

THE DIRECT DATA PLATFORM ™

incorta

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