Loading Data

JN **John Nadasky**

Extraction and Loading Process

It is important to understand that Loading Data represents a process that consists of two phases: **Extraction** and **Load**. A Load Job manages both phases of the Loading Data process.

The first phases of the processes is **Extraction**, which the **Loader Service** manages:

01 | Loader Service

Extract the data for schema tables

The Loader Service extracting data from the data source as defined by the physical schema table. Extraction may mean reading a local file, executing a SQL query, or consuming a Kafka topic.

02 | Loader Service

Create Apache Parquet files in Shared Storage

The Loader Service creates Apache Parquet files in Shared Storage (Staging) for the extracted records sets.

03 | Loader service

In-Memory Direct Data Map

The Loader Service reads the Apache Parquet files and creates an in-memory Direct Data Map of the key indexes and joins between tables in the schema.

04 | Loader Service

Create Direct Data Mapping (DDM) files in Shared Storage

The Loader Service persists the Direct Data Mapping schema to Shared Storage (Staging) as Direct Data Mapping files (snapshots).

The second phase of the processes is to **Load** the Direct Data Map for the schema into the **Analytics Service**:

05 | Analytics Service

Reads the Direct Data Mapping files into memory

The Analytics Service reads the Direct Data Mapping files from Shared Storage into memory.

06 | Analytics Service

Reads the Apache Parquet files into memory

The Analytics Service reads the Apache Parquet files from Shared Storage for the schema tables and loads the table data in compressed form into memory.

Loading Data

In Incorta, administrators or developers simply load the data for a schema or a table in a schema. Loading typically encompasses both the extraction by the Loader Service and the loading into memory for the Analytics Service.

For a schema load, you have the option to load a schema on demand or schedule when a schema loads. For a table load, you only have the option to load data on demand.

For loading a schema on demand, there are three options:

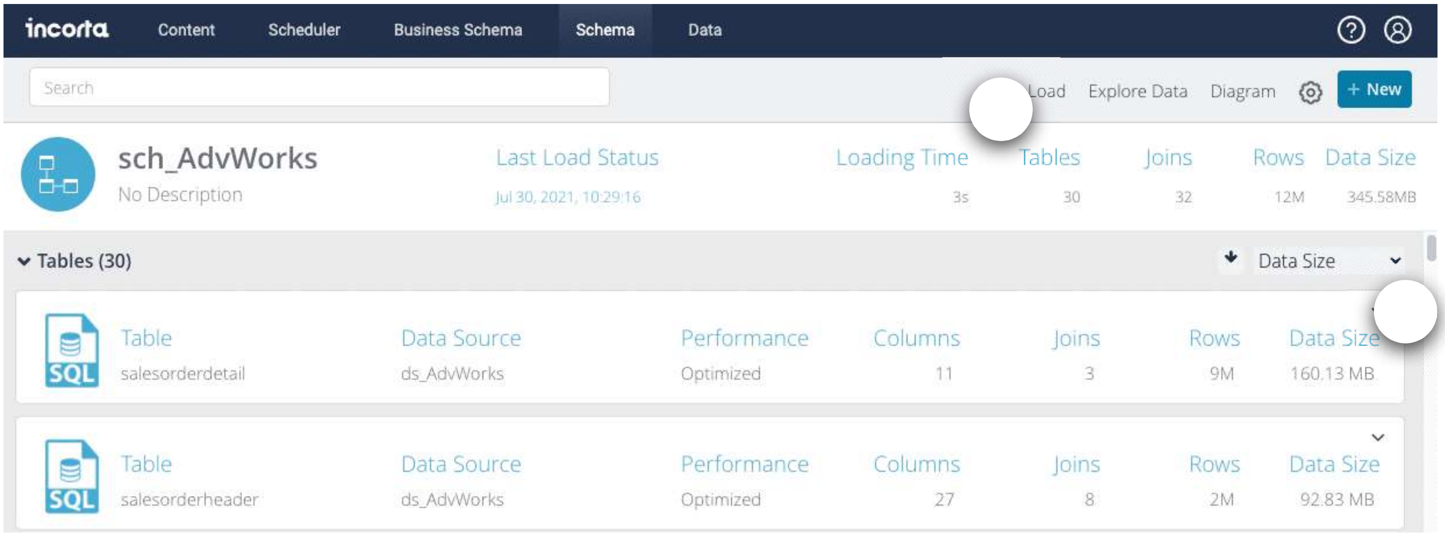
- Full
- Incremental
- Staging

For loading on a table on demand, there are two options:

- Load table (Full)
- Load from Staging

> Incorta locks the schema during a load. During a Load job, you cannot modify a schema.

Explore the following infographic to show how you can load a given schema and load given table.



sch_AdvWorks
No Description

Last Load Status	Loading Time	Tables	Joins	Rows	Data Size
Jul 30, 2021, 10:29:16	3s	30	32	12M	345.58MB

Tables (30) — Data Size

Table	Data Source	Performance	Columns	Joins	Rows	Data Size
salesorderdetail	ds_AdvWorks	Optimized	11	3	9M	160.13 MB
salesorderheader	ds_AdvWorks	Optimized	27	8	2M	92.83 MB

Full

The Loader Service extracts data from the data source for the schema table or for the single specified table. Incorta stores the extracted data as Apache Parquet files in Shared Storage (Staging). Incorta then uses the Parquet file data to create internal indexes and join paths in-memory as a Direct Data Map.

Incorta persists the Direct Data Map to Shared Storage as Direct Data Mapping files. For tables with Performance Optimized enabled, the Analytics Service will load into memory the related Direct Data Mapping files from Shared Storage.

Incremental

Generally, the Incremental option applies only to loading tables that are configured as incremental tables. A table enabled for incremental loads typically contains a primary key and timestamp column.

There is on exception to this: an Apache Kafka table.

The **Incorta 4 Foundations for Developer** course covers advanced data loading strategies and tactics such as configuring Incremental tables and scheduling incremental loads. The course also covers how to ingest data from Apache Kafka.

To learn more about his course, visit learn.incorta.com.

For an incremental table, you can specify an update option, either a SQL select statement or an update file. A configuration setting for incremental tables can exclude the table from a Full load. An Incremental load applies to all incremental tables in a schema.

Incorta queries the incremental data from the data source or data file and in turn stores the extracted data as Apache Parquet files in Shared Storage (Staging). Incorta then uses the Parquet file data to create and update internal indexes and join paths in-memory as a Direct Data Map for the incremental tables. Incorta persists the Direct Data Map to Shared Storage as Direct Data Mapping files. For tables with Performance Optimized enabled, the Analytics Service will load into memory the related Direct Data Mapping files from Shared Storage as well as load data from Parquet files.

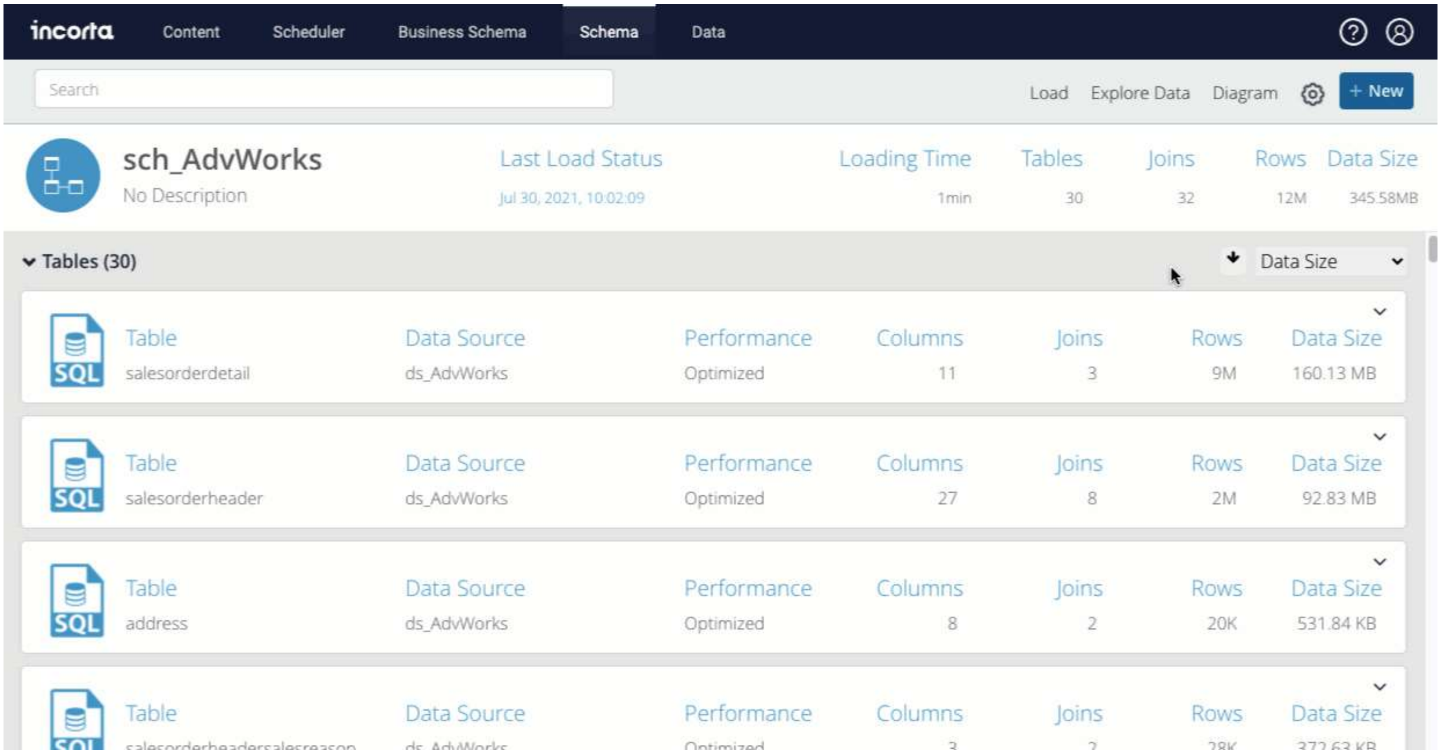
Because schemas can change between incremental loads, Incorta invokes a compaction process after every incremental load. This background process provides a consistent version of the incremental table's data. Consistency in this regard means a consistent number of columns and consistent data.

Performance Optimized

By default, Incorta enables Performance Optimization for all tables in a schema. The Analytics Service will only load tables that are Performance Optimized. In other words, if Performance Optimized is disabled for a given table, it will not be loaded into the Analytics Service.

You can enable or disable Performance Optimized setting on an individual table in the Table Editor. You can also enable and disable the Performance Optimized setting for one or more tables in Schema Settings.

The following animation details how to enable and disable performance optimization for all tables in a given schema as well as enable performance optimization for a single table. Notice the warning about data eviction from the Analytics service and the recommendation to load from staging.



sch_AdvWorks
No Description

Last Load Status	Loading Time	Tables	Joins	Rows	Data Size
Jul 30, 2021, 10:02:09	1min	30	32	12M	345.58MB

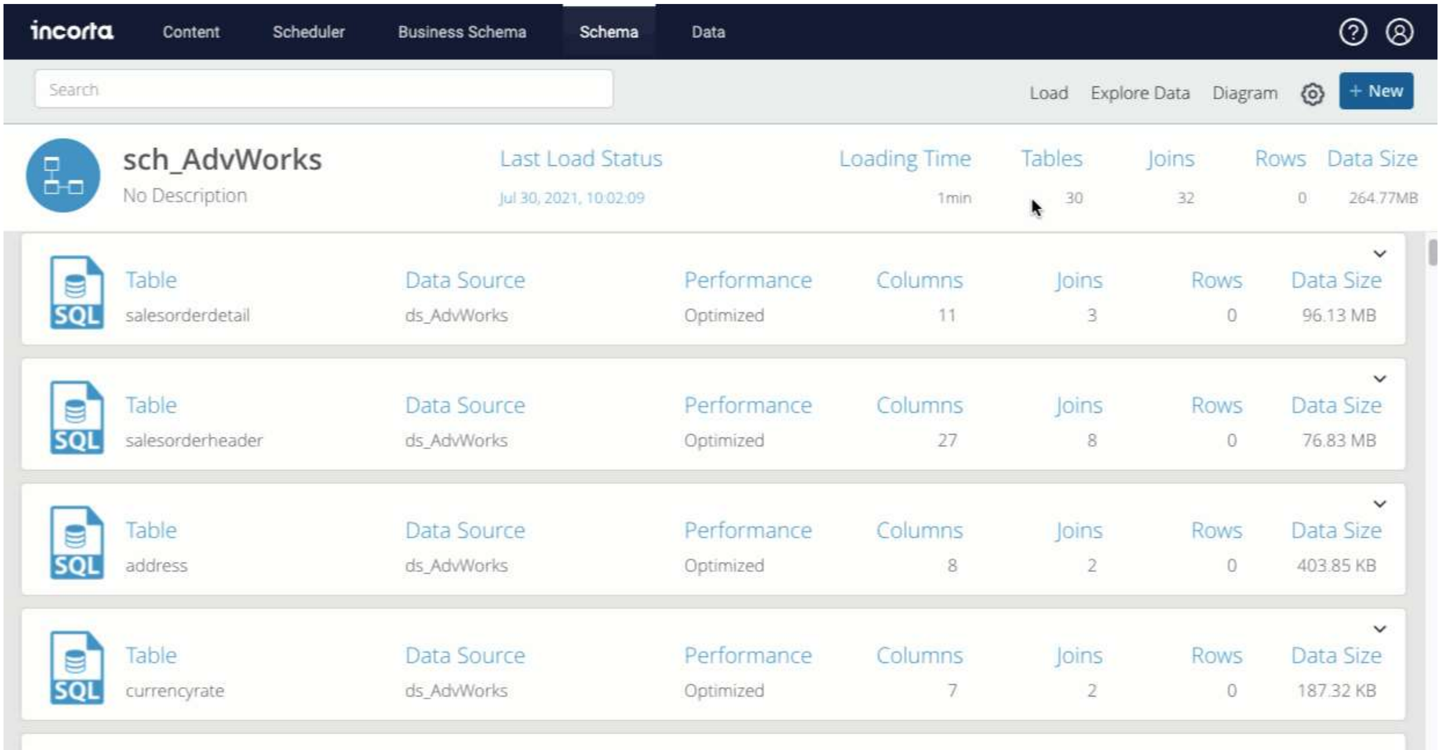
Table	Data Source	Performance	Columns	Joins	Rows	Data Size
salesorderdetail	ds_AdvWorks	Optimized	11	3	9M	160.13 MB
salesorderheader	ds_AdvWorks	Optimized	27	8	2M	92.83 MB
address	ds_AdvWorks	Optimized	8	2	20K	531.84 KB
salesorderheadersalesreason	ds_AdvWorks	Optimized	3	2	28K	372.63 KB

Staging (load from Shared Storage)

The Staging load option instructs the Loader Service to **not extract data** from the data source for the schema table or for the single specified table.

Instead, the Analytics Service will load into its in-memory Direct Data Mapping engine the related Direct Data Mapping and Parquet files from Shared Storage (Staging).

The following animation details there are no rows extracted, only rows loaded.



sch_AdvWorks
No Description

Last Load Status	Loading Time	Tables	Joins	Rows	Data Size
Jul 30, 2021, 10:02:09	1min	30	32	0	264.77MB

Table	Data Source	Performance	Columns	Joins	Rows	Data Size
salesorderdetail	ds_AdvWorks	Optimized	11	3	0	96.13 MB
salesorderheader	ds_AdvWorks	Optimized	27	8	0	76.83 MB
address	ds_AdvWorks	Optimized	8	2	0	403.85 KB
currencyrate	ds_AdvWorks	Optimized	7	2	0	187.32 KB

That's all folks.