DBImport

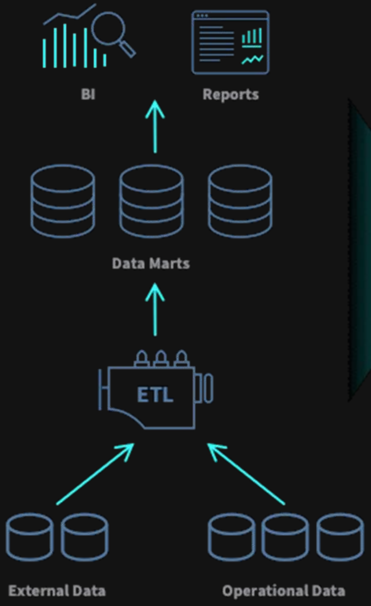
The smart way to ingest data into a Data Lakehouse

Middlecon

Evolution of data loading for analytics

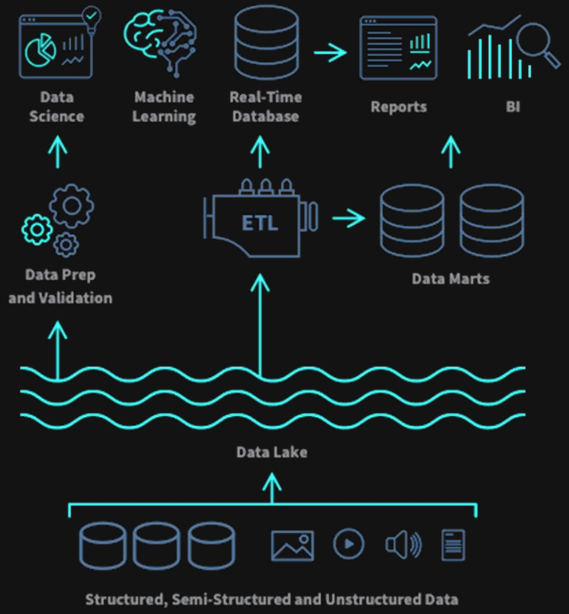
LATE 1980'S

Data Warehouse



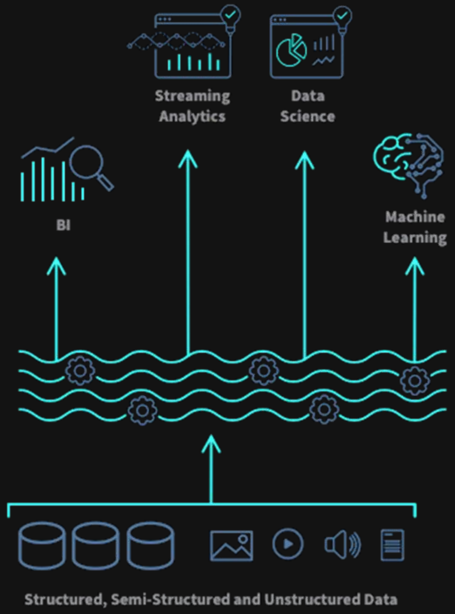
2011

Data Lake



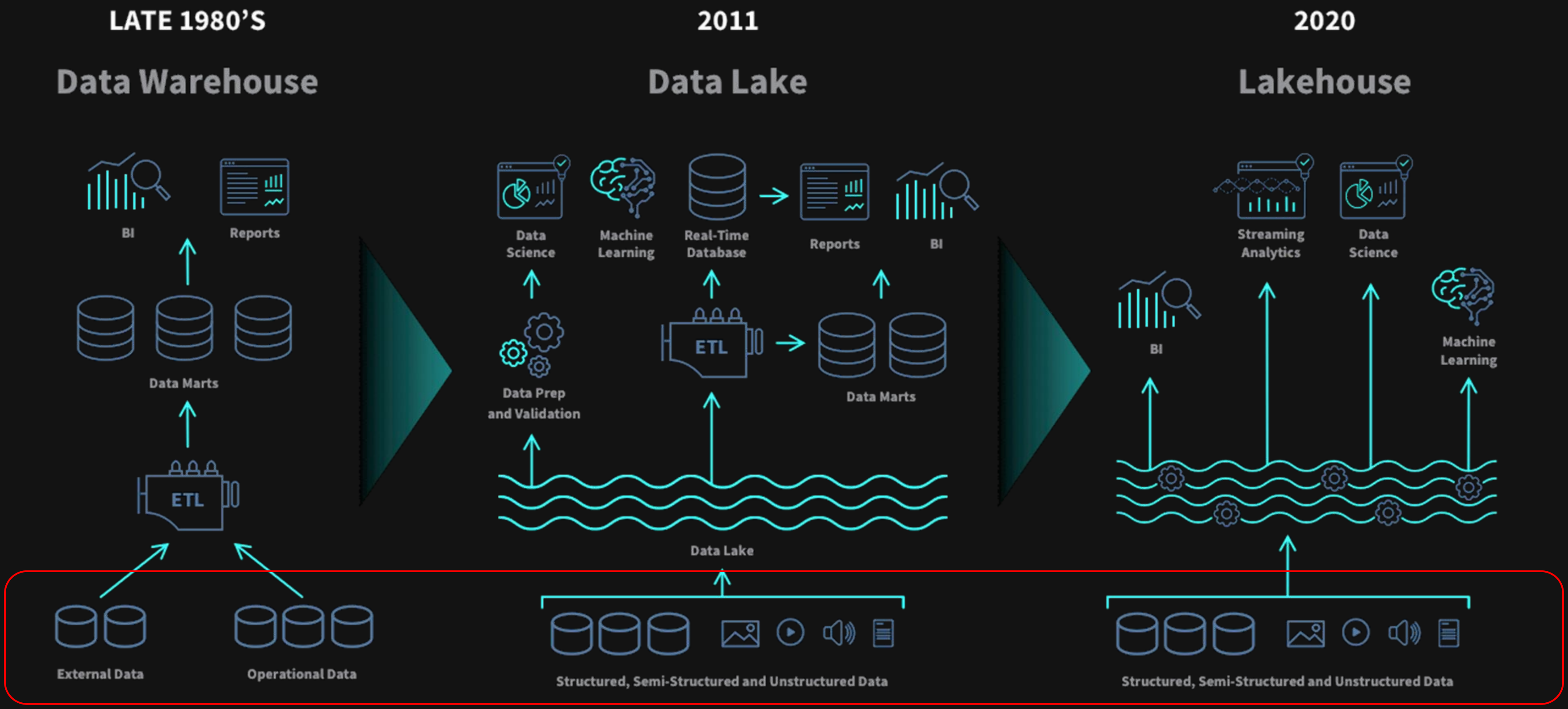
2020

Lakehouse



Middlecon

Evolution of data loading for analytics



Middlecon

Data Ingestion

Regardless of data evolution, we will always have to move and load data into a system

Data Ingestion Challenges

- Large tables take forever to ingest
- Handling incremental data synchronization
- Source system owners says no to CDC tools
- Schema Changes in the source system

Data Ingestion Wish-List

- Support for all common database types out of the box
- Auto discovery of tables and views from source systems
- Full and incremental import and export functionality
- Automatic handling of schema changes
- Automatic handling of table and column descriptions
- Automatic handling of primary and foreign keys

Data Ingestion Wish-List

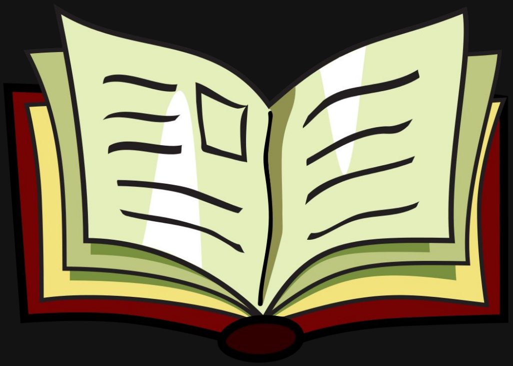
- Change Data Tracking functions
- Audit of changed metadata
- Parallel execution and central scheduling
- Full logging and statistics of data ingestion processes
- Modern fileformat usable by many tools
- Ease of use

DBImport

Key functionalities

DBImport Tool - History

Many years of experience with different sqoop, spark and jdbc data ingestions problems resulted in an OpenSource project called DBImport



Largest installation are ingesting data from over 19.000 tables in a single Hadoop environment per day, from 250 different sources

Licensed by Apache 2.0

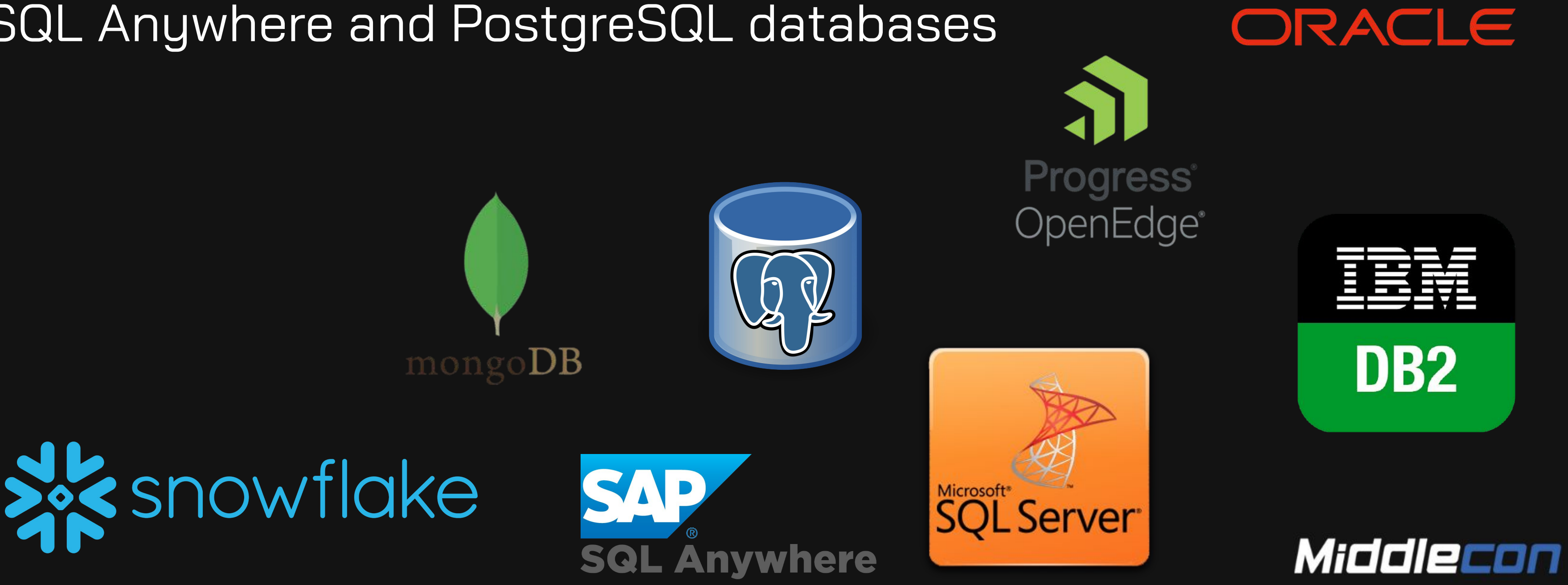
DBImport Tool - Goal

The goal of DBImport is to have a fast, simple but powerful tool to create a source aligned copy of the data and at the same time keeping track of what data have changed between two different points in time.

The target is always a Data Lake or Data Lakehouse and the source aligned data storage is usable for all AI, Machine learning, BI and exploration of data within the organization.

DBImport Tool - Supported source systems

Support for Oracle, MsSQL, MySQL, DB2 UDB, DB2 AS400, MongoDB, Snowflake, Progress, SQL Anywhere and PostgreSQL databases



DBImport - Extract methods

Both Full extraction and Incremental extraction is supported through standard SQL.

Additionally, both MSSQL Change Data Tracking and Oracle Flashback Query can be used to extract data from source system

ORACLE

Microsoft SQL Server

DBImport - Extract tool

Uses sqoop or spark in the background for transferring data from source system

DBImport - Fileformats and Storage

DBImport will save data in Orc, Parquer or Iceberg format on either HDFS or Ozone.

Apache orc™

Parquet

ICEBERG

hadoop

DBImport - Load and transform tool

Once data is loaded from source system, transformation is handled by Spark or Hive

DBImport - Data access

Data accessible as files on HDFS or Ozone or through Hive and Impala with standard SQL

Apache Impala

Apache orc™

Parquet

HIVE

ICEBERG

DBImport Tool - Supported export systems

Exports data stored in Hive.

Support for Oracle, MsSQL, MySQL, DB2 UDB, DB2 AS400, MongoDB, Snowflake, Progress, SQL Anywhere and PostgreSQL databases. Also support for creating files on AWS S3

DBImport - History Data

If selected, DBImport will keep track of what data has been changed in a source system table and log this in a separate history table.

This history table can then be used to incrementally process data even if the source system does not support incremental data loads or keep track of changed data.

It can also be used to see how data is changed over time.

Id	Type	Stock	datalake_iud	datalake_timestamp
1	apple	205	I	2019-09-08 02:18:45
2	orange	155	I	2019-09-08 02:18:45
3	banana	40	I	2019-09-08 02:18:45
4	pear	70	I	2019-09-08 02:18:45
5	kiwi	65	I	2019-09-08 02:18:45
1	apple	200	U	2019-09-09 02:20:40
2	orange	125	U	2019-09-09 02:20:40
5	kiwi	65	D	2019-09-09 02:20:40

Middlecon

DBImport - Auto Discovery and source system changes

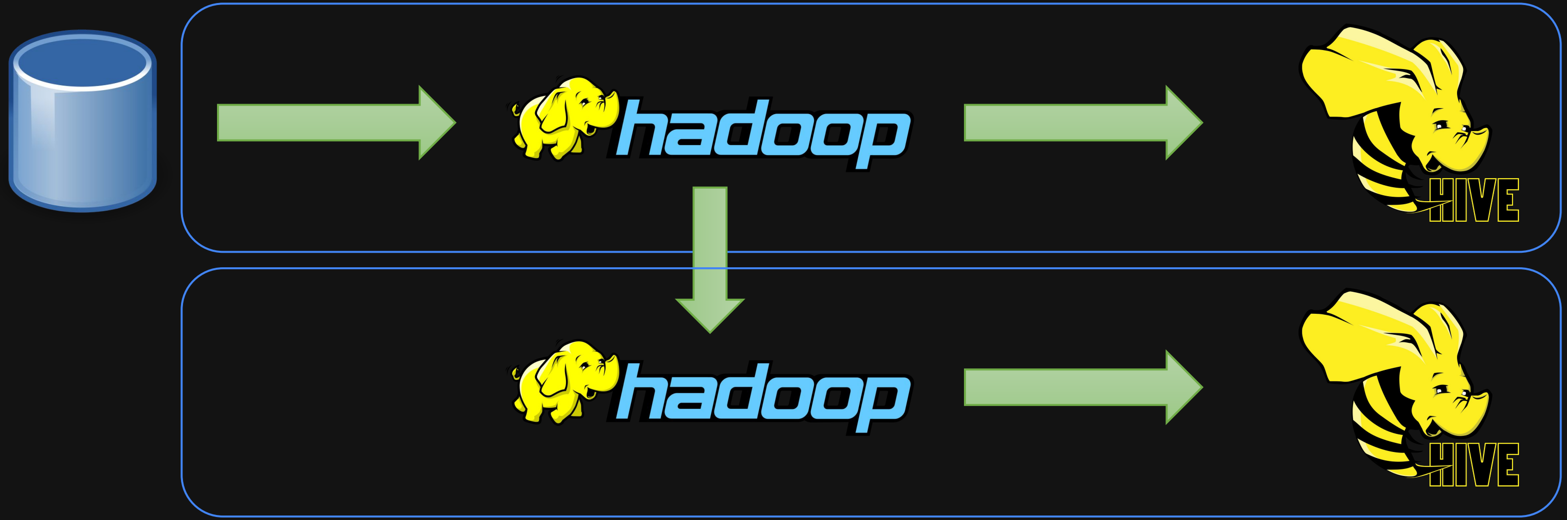
Auto discovery of tables and views from all supported source systems.

All changes on column and table changes including keys and comments is identified and changed in target tables. All changes is logged and accessible by third-party tools.

Middlecon

DBImport Tool - Multi Cluster Ingestions

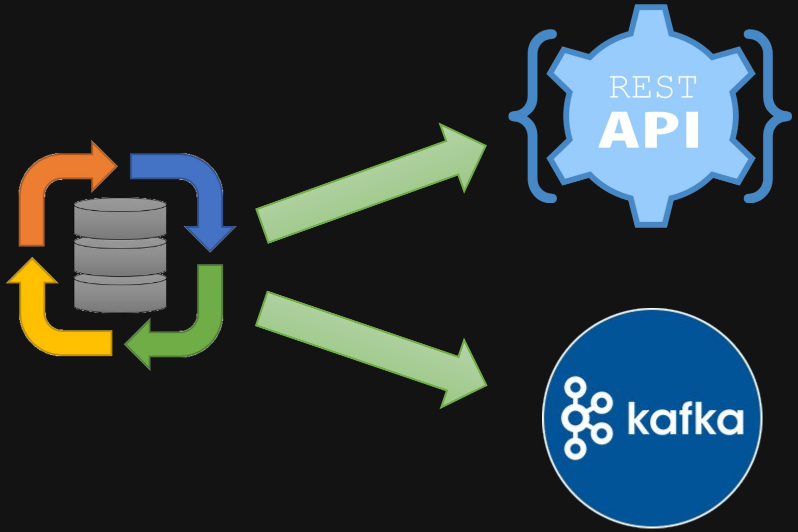
Ability to keep two system updated with only one extraction from source system



Middlecon

DBImport Tool - Notification

After each task is completed, a notification will be posted through a Rest API or on a Kafka topic.
JSON Data includes statistics and status of the job



DBImport Tool - Statistics

All completed tasks will store statistics about how the task was executed. Both time and the amount of data that was moved.

DBImport Tool - Verifications and Restarts

During the import, there is verifications three times on the imported data to make sure that the data that was fetched is the same as on the source system.

If the data is not correct, or due to other technical reason the job failed the task will be restarted from last successful state

DBImport Tool - Data Anonymization

During transport of data from source system, data can be anonymized before it is stored on any permanent storage device.

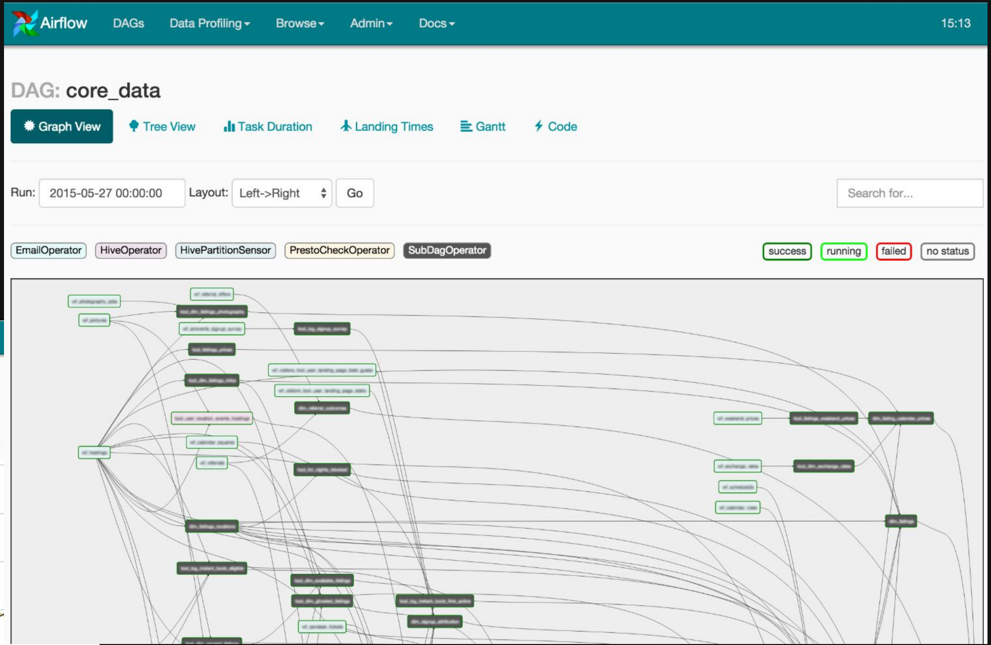
Supports Hash, ‘Replace with star’ and ‘Show first 4 chars only’ methods. Selectable on column level

DBImport Tool - Airflow

Apache
Airflow

DBImport Tool - Airflow Integration

Airflow is an OpenSource scheduling and Workflow management system



Airflow DAGs Data Profiling▾ Browse▾ Admin▾ Docs▾ 15:13

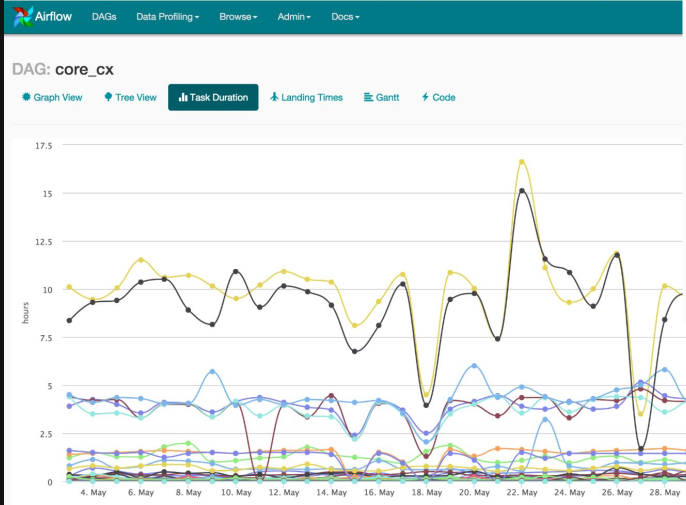
DAG: core_data

Graph View | Tree View | Task Duration | Landing Times | Gantt | Code

Run: 2015-05-27 00:00:00 Layout: Left->Right Go Search for...

EmailOperator HiveOperator HivePartitionSensor PrestoCheckOperator SubDagOperator

success running failed no status



Airflow DAGs Data Profiling▾ Browse▾ Admin▾ Docs▾

DAG: core_cx

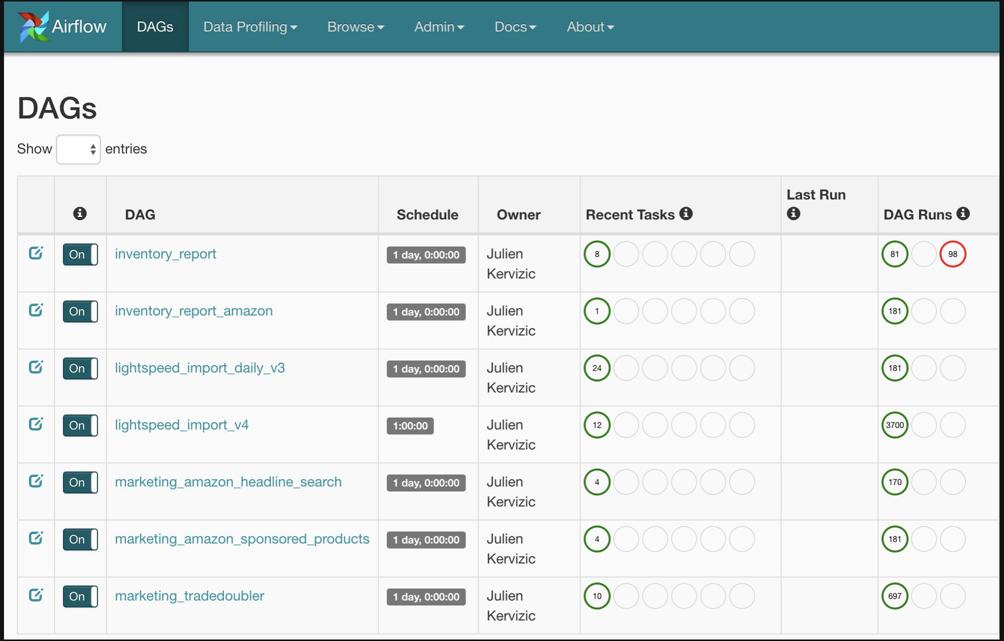
Graph View | Tree View | Task Duration | Landing Times | Gantt | Code

Middlecon

DBImport Tool - Airflow Integration

DBImport generates Airflow DAG's and writes them to the DAG directory of Airflow.

It also generates pools in Airflows configuration database, one per source database hostname and one per DAG



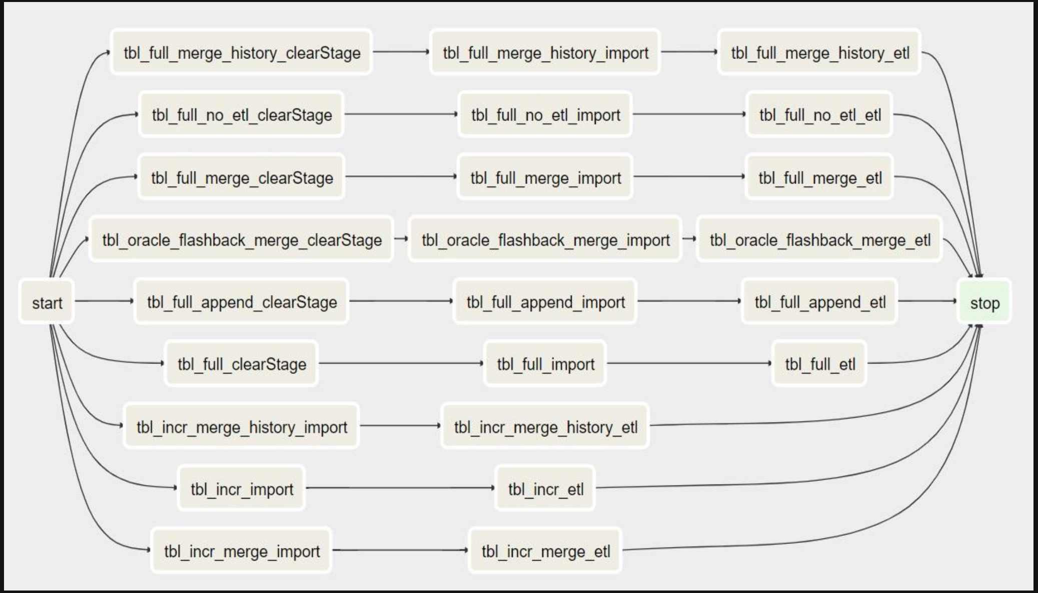
DAGs

Show entries

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs
	On	inventory_report	1 day, 0:00:00	Julien Kervizic	8		81, 98
	On	inventory_report_amazon	1 day, 0:00:00	Julien Kervizic	1		181
	On	lightspeed_import_daily_v3	1 day, 0:00:00	Julien Kervizic	24		181
	On	lightspeed_import_v4	1:00:00	Julien Kervizic	12		3700
	On	marketing_amazon_headline_search	1 day, 0:00:00	Julien Kervizic	4		170
	On	marketing_amazon_sponsored_products	1 day, 0:00:00	Julien Kervizic	4		181
	On	marketing_tradedoubler	1 day, 0:00:00	Julien Kervizic	10		697

DBImport Tool - Airflow Integration

Each DAG contains the required steps for all tables being imported in the DAG. This usually is all the tables from the source database that is being imported.

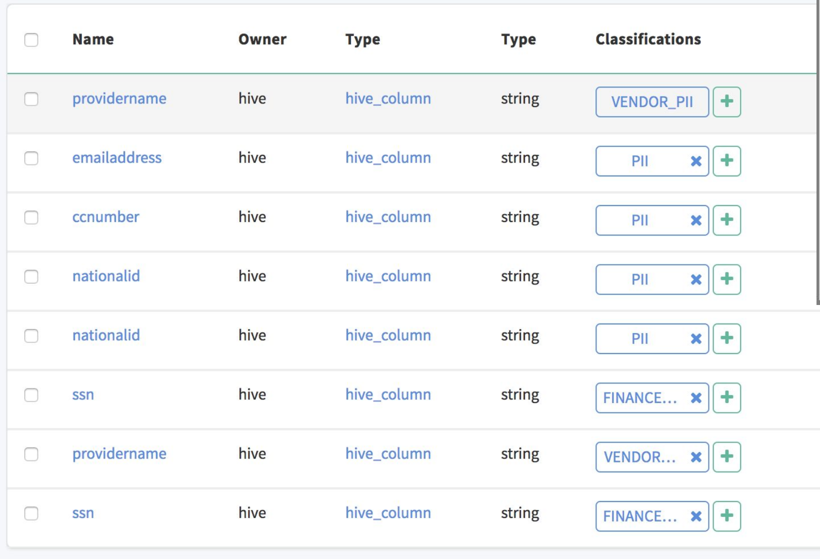

start
tbl_full_merge_history_clearStage
tbl_full_merge_history_import
tbl_full_merge_history_etl
tbl_full_no_etl_clearStage
tbl_full_no_etl_import
tbl_full_no_etl_etl
tbl_full_merge_clearStage
tbl_full_merge_import
tbl_full_merge_etl
tbl_oracle_flashback_merge_clearStage
tbl_oracle_flashback_merge_import
tbl_oracle_flashback_merge_etl
tbl_full_append_clearStage
tbl_full_append_import
tbl_full_append_etl
tbl_full_clearStage
tbl_full_import
tbl_full_etl
tbl_incr_merge_history_import
tbl_incr_merge_history_etl
tbl_incr_import
tbl_incr_etl
tbl_incr_merge_import
tbl_incr_merge_etl
stop

DBImport Tool - Atlas

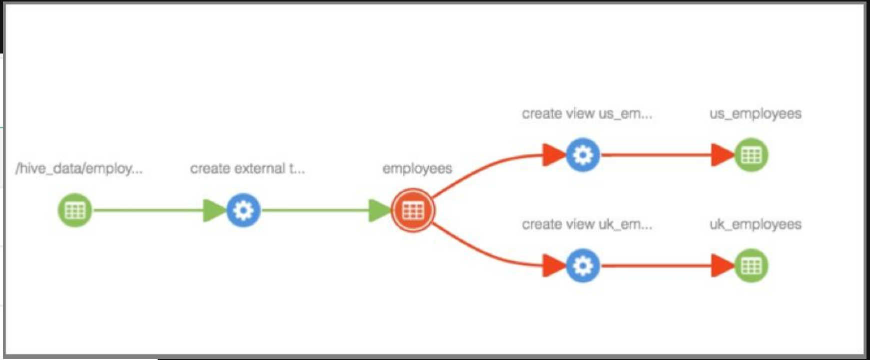
Apache Atlas

DBImport Tool - Atlas

Apache Atlas is an open-source metadata and big data governance framework

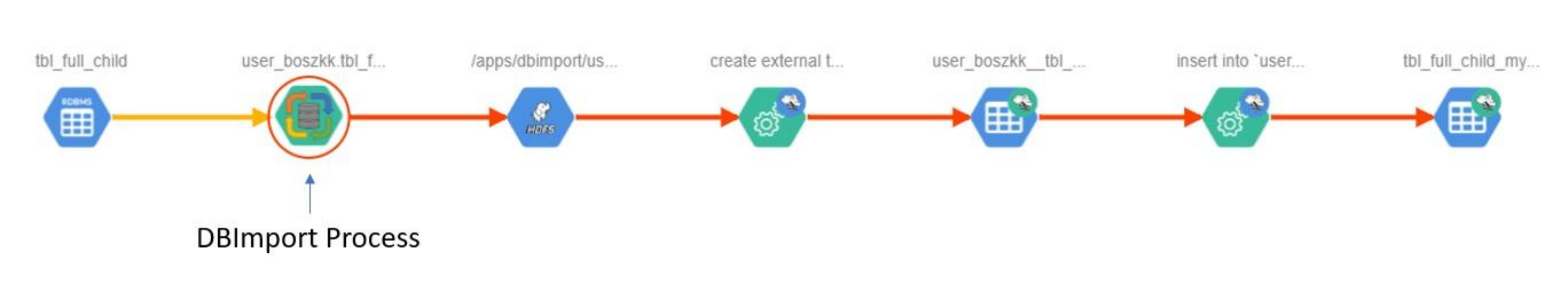


☐	Name	Owner	Type	Type	Classifications	
☐	providername	hive	hive_column	string	VENDOR_PII +	
☐	emailaddress	hive	hive_column	string	PII ✖ +	
☐	ccnumber	hive	hive_column	string	PII ✖ +	
☐	nationalid	hive	hive_column	string	PII ✖ +	
☐	nationalid	hive	hive_column	string	PII ✖ +	us_customers
☐	ssn	hive	hive_column	string	FINANCE... ✖ +	tax_2015
☐	providername	hive	hive_column	string	VENDOR... ✖ +	provider_summary
☐	ssn	hive	hive_column	string	FINANCE... ✖ +	tax_2010



DBImport Tool - Atlas Integration

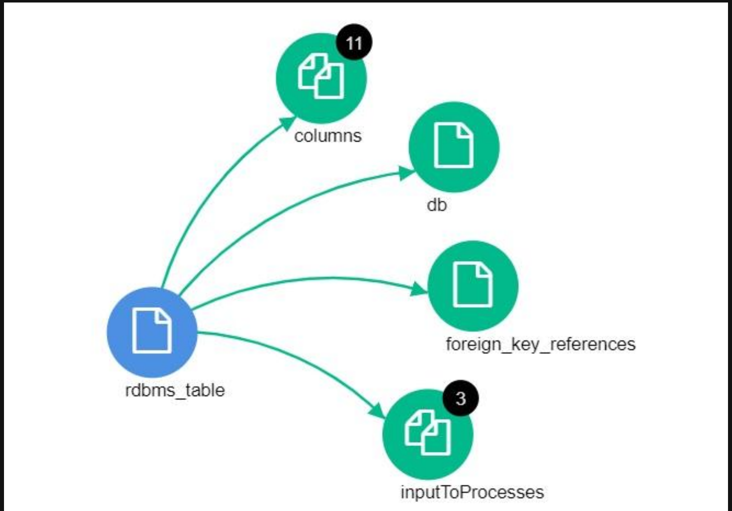
Full lineage is available with the help of the custom DBImport Process Atlas type.



DBImport Process

DBImport Tool - Atlas Integration

DBImport utilizes the rdbms_* types already existing in Atlas to create a complete source system object. Source system object is complete with all information and relationship for the imported / exported table



DBImport Tool - Atlas Integration

Source system information can be added in two steps.

- When an import or export is executed, the information for that specific table is created / updated in Atlas.
- When the DBImport server is discovering ALL tables on a specific connection, regardless if it's imported or not to Hadoop

DBImport

Customer Use-Case

DBImport Tool - Customer Use Case

Large Swedish company in manufacturing industry uses DBImport for close to all batch ingestion into their Datalake / Data lakehouse.

- 19.000 tables per day
- 800 GB per day
- 41 billion rows per day
- 450 Airflow jobs

Total imported size over time is 1.1 PB with DBImport

DBImport

Current and Future development
and improvements

DBImport Tool - Future

- Python based Windows Client
  In Development
- Spark as ETL engine
  In Development - Beta version available
- Container based cloud application
  On Road-Map

Questions?

Like to contribute?

Please contact Berry Österlund

DBImport

Git

https://github.com/Middlecon/DBImport

Documentation

https://dbimport.readthedocs.io/en/latest/

Lead developer

berry.osterlund@middlecon.se