TETMON

SOLUTION PAPER

Unique Features & Algorithms Behind EdgeSet – 2022

contact@tetmon.com | www.tetmon.com

CONTENT

EdgeSet joins disparate cloud and on-premise databases to a single point of access, and outputs to your favourite BI visualisation tools

PROBLEM

Decision-makers need data to make informed decisions, but data spread across disparate databases

Building central data warehouse takes 1-2 years, $1-10M, 10-20 engineers; maintaining ETL is painful, and breaks after weeks

SOLUTION

Virtual central data warehouse in minutes; both on-premise and web deployment available

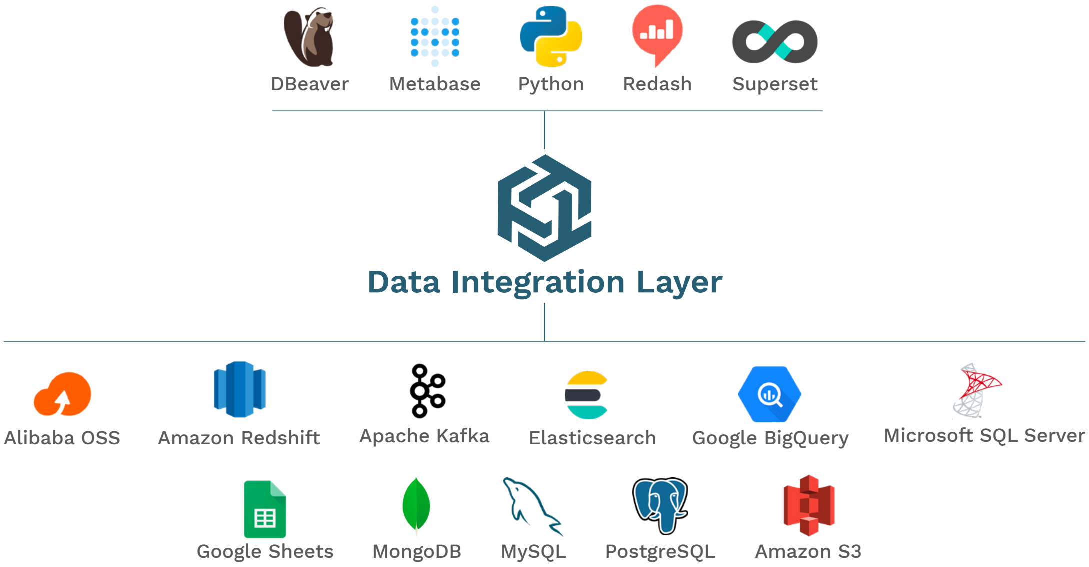
Data transforms now cheap & easy; reducing ETL & required data engineering headcount

KEY FEATURES

DATA DISCOVERY
EdgeSet removes the need for Hive/ AWS Glue. 4 weeks down to 40 minutes

Automatic schema inference for non-relational data sources, including Amazon S3 and Google Sheets

EASE OF USE
Trino deployment from minimum 2 weeks of Data Engineering time, down to 2 hours



Features and Algorithms Unique to EdgeSet

EdgeSet contains some unique algorithms that have never before been applied to data scanning and data warehousing.

1. Automatic Table Inference

EdgeSet is the first software to support importing semi-structured data (i.e. tabular files) without any manual data engineering work. To illustrate this, we'll compare the process for importing CSV files from an Amazon S3 bucket into Snowflake, Apache Hive, and EdgeSet:

SNOWFLAKE

1. Manually examine all of the files.
2. Group files into tables, noting down all of the file names.
3. For each table:
   3.1. Determine the file format of the associated files, the column names, and the types of the columns.
   3.2. Write a `CREATE TABLE` statement using the column names and column types.
   3.3. Write a regular expression that matches all associated file names.
   3.4. Write a `COPY INTO` statement to copy the data from the file into the target table using the previously noted file format.

VS

APACHE HIVE

1. Manually examine all of the files.
2. Group files into tables, noting down all of the file names.
3. If table partitions don't fit Hive's expected format, rename all the files.
4. For each table:
   4.1. Determine the file format of the associated files, the column names, and the types of the columns.
   4.2. Write a `CREATE TABLE` statement using the column names, column types, file format, and partition format.
   4.3. Test that the table is usable by querying it.
   4.4. Repeat step 4.2 adjusting the `CREATE TABLE` statement using trial and error until the table gives expected results.

VS

EDGESET

1. Enter the AWS S3 bucket name and API key.

EdgeSet will scan the S3 bucket, infer which files should be grouped into tables, infer the column names and types, and create the necessary tables automatically.

Why is automatic table inference important?

It's not just tedious work to have to write up a `CREATE TABLE` statement for each (set of) tabular file(s) that you want to import—it's also error-prone. And the errors introduced this way are often silently ignored by data warehouse systems, leading to corrupt data and misguided analysis that can persist undetected indefinitely.

Some tools provide a guided file import, showing a preview of the file contents and allowing for quick visual feedback while selecting column names and data mappings. While this guided approach is more convenient than Snowflake and Apache Hive processes, it is still prone to errors and does not scale to large numbers of files and/or tables.

How does EdgeSet perform automatic table inference?

EdgeSet uses a collection of custom parsers whose output types form a semilattice. These parsers allow EdgeSet to detect dates and times in multiple formats and languages (a wider variety of formats than those found in date parsing libraries), arbitrary-precision decimal numbers, and various common industry-specific data formats. The semilattice allows EdgeSet to do whole-table inference, combining hints from multiple rows across multiple files to resolve ambiguities and ensure that no values are silently truncated or ignored.

2. Spreadsheet Data Range Detection

EdgeSet is the only software that supports fully-automatic Google Sheets import, including Google Drive scanning and in-sheet data range (table) detection.

Why is automated data range detection important?

Without automated data range (table) detection, the table must be formed over the entire spreadsheet range or the target data range must be specified manually. However, spreadsheets often contain marginal annotations and summary calculations (formulas). If these non-data cells are included in the table, they can lead to incorrect query results and aggregate statistics.

How does EdgeSet detect data tables?

EdgeSet scans each spreadsheet to detect all cell ranges that could form possible data tables and performs table inference on all such ranges. Using feature coefficients from the scan and inference steps, EdgeSet then narrows the tables to the best candidates using a model that was pre-trained on a wide variety of real life spreadsheets. The overall process is similar to how machine vision systems scan images and recognize objects (and in fact EdgeSet uses some of the same algorithms).

3. On-demand Row Fetching and Dynamic Query Keep-alive

EdgeSet's built-in web interface uses server-sent events to keep the user informed of a query's progress the moment anything changes. A custom nested virtual scrolling algorithm allows EdgeSet's UI to calculate just when and how many records to request from the backend in order to keep the UI responsive and seamless without requesting unnecessary work from the connected data sources. This gives the user the same data exploration experience whether exploring hundreds or trillions of records.

CASE STUDY 1

Joining multiple Excels at a premier financial institution

PROBLEM

Operations reconciliation tedious and error-prone

EDGESET AS A SOLUTION

Use EdgeSet to join multiple excels across different departments

BENEFITS

- Back-office operations staff now able to go home on-time instead of midnight
- For a four-man team, 6 man-hours saved, per day; with a reduction in staff resignations
- Errors during trade reconciliation identified quickly, thereby reducing amplified errors to middle office and front office

SITUATION BEFORE TETMON

- 4-person Operations team has to regularly work till midnight to reconcile trades
- Data errors hinder alpha generation

AFTER IMPLEMENTING EDGESET

- Operations staff can now go home on-time
- Automation reduced ‘eye-ball’ checks

COST BENEFITS

- 6 manhours saved, per day

OTHER BENEFITS

- Discover market inefficiencies

Next Stage:
Explore joining Excels with internal data sources

CASE STUDY 2

Normalizing currencies for a multinational's data across 8 countries

MOTIVATION

- Provide summary statistics across operations in 8 countries in EUR
- Compare customer lifetime value, acquisition costs, and profitability across countries

PRIOR DATA SITUATION

- Financial data in 8 different MySQL databases
- Reports generated individually for each country
- Analysts had to then manually roll up the reports in Excel and convert currencies using the preferred method of the analyst

PROBLEM

Difficulty in obtaining summary operation statistics across 8 countries' operations; analysts in 8 countries had difficulty keeping in sync with headquarter's approved foreign exchange rates

EDGESET AS A SOLUTION

Headquarters now inputs approved currency conversion rates into a shared Google Sheet. Analysts in 8 countries use EdgeSet to connect to the spreadsheet and all 8 countries' databases

BENEFITS

- Efficiently compare customer lifetime value, acquisition costs, and profitability across countries
- Reduced human error during manual excel rollups office and front office

COMPARISON

	WITHOUT EDGESET	WITH EDGESET
APPROACH	HQ regularly emails approved currency conversion rates to each country. Each country's engineering team inputs the rates into the database	HQ regularly inputs approved currency conversion rates into a shared Google Sheet. Analysts use EdgeSet to connect to the spreadsheet + all 8 country's databases
SETUP TIME	None	4 hours to install EdgeSet and setup analyst permissions
CONVERSION RATE INPUT TIME	1 hour x 9 per week	1 hour per week
CONVERSION RATE LAG	Varies	-
ROLLUP REPORT PREPARATION	1 day	1 hour
LIMITATIONS	Manual Excel rollups prone to human error	-

CASE STUDY 3

Alternative to expensive Oracle license fee and server upgrade costs, for a major online retailer

MOTIVATION

- Link transactions to website visits and marketing campaigns for accurate attribution
- Remove poor-performing channels from marketing spend and invest more in top-performing channels

PRIOR DATA SITUATION

- Transactional data loaded to Oracle data warehouse via Pentaho
- Click data in Google Analytics Premium (not in data warehouse)
- Google Analytics data was not used by analysts except through the web interface

PROBLEM

A major online retailer would like to link transactions to website visits and marketing campaigns for accurate attribution. In addition, remove poor-performing channels from marketing spend and invest more in top- performing channels. However, Oracle license fee and server upgrade costs are prohibitively expensive

EDGESET AS A SOLUTION

Direct connection to Oracle and Google Analytics via BigQuery (No ETL) by using EdgeSet

BENEFITS

- Cost savings of $300,000
- Reduce ETL Running time and lag time
- Full click history and data maintained without the need for constant deletion

COMPARISON

	WITHOUT EDGESET	WITH EDGESET
APPROACH	Add click data to ETL pipeline via BigQuery and import to data warehouse	Direct connection to Oracle and Google Analytics via BigQuery (No ETL)
SETUP TIME	2 weeks to setup ETL graphs in Pentaho + 1 week to migrate to larger Oracle server	4 hours to install EdgeSet and migrate permissions from Oracle
ETL RUNNING TIME	30 minutes for yesterday's data	None
ETL LAG TIME	24 hours	None
SQL QUERY WRITING EFFORT	Same	Same
REPORT GENERATION	1-2 minutes	3-5 minutes
OTHER COSTS	$282,000 additional Oracle license fee; $16,000 server upgrade	Per-query BigQuery charges (varies, but ~$1/query)
LIMITATIONS	Oracle data warehouse not capable of maintaining full click history and old data must be deleted	None

REFERENCES & AWARDS

- Established software system integrators distribute EdgeSet in Asia
- Singapore's representative to Chongqing Tech Unicorn Summit 2020, International Dedicated Connectivity Forum 2021, Speaker at Chief Data Officer Live 2021

PRINCIPALS

HU YINGHAN
CEO

Yinghan was General Manager for a Singapore-Chongqing Cross-border Connectivity Initiative; held management and strategy positions at Singtel, SPH and started his career as a GIC Investment Associate. He volunteers on Singapore Fintech Association's ESG sub-committee, is a member of Ant Financial's inaugural 10x1000 Fintech Leaders batch and is Local CONNECT Lead for 10x1000, a philanthropic initiative launched jointly by the International Finance Corporation (IFC) and Alipay. Yinghan graduated from Oxford University with a BA and MA in Jurisprudence.

CHRISTOPHER FORNO
CTO (GITHUB.COM/JEKOR)

For two decades, Chris has built and led global software engineering teams to execute challenging projects ranging from very large distributed systems to computer vision, at employers/ clients including Rocket Internet, Lazada, Yucaipa, DeviantArt. His youtube programming tutorials has been viewed >500k times