Experiment 10

Ojas Patil
2019130048
TE COMPS

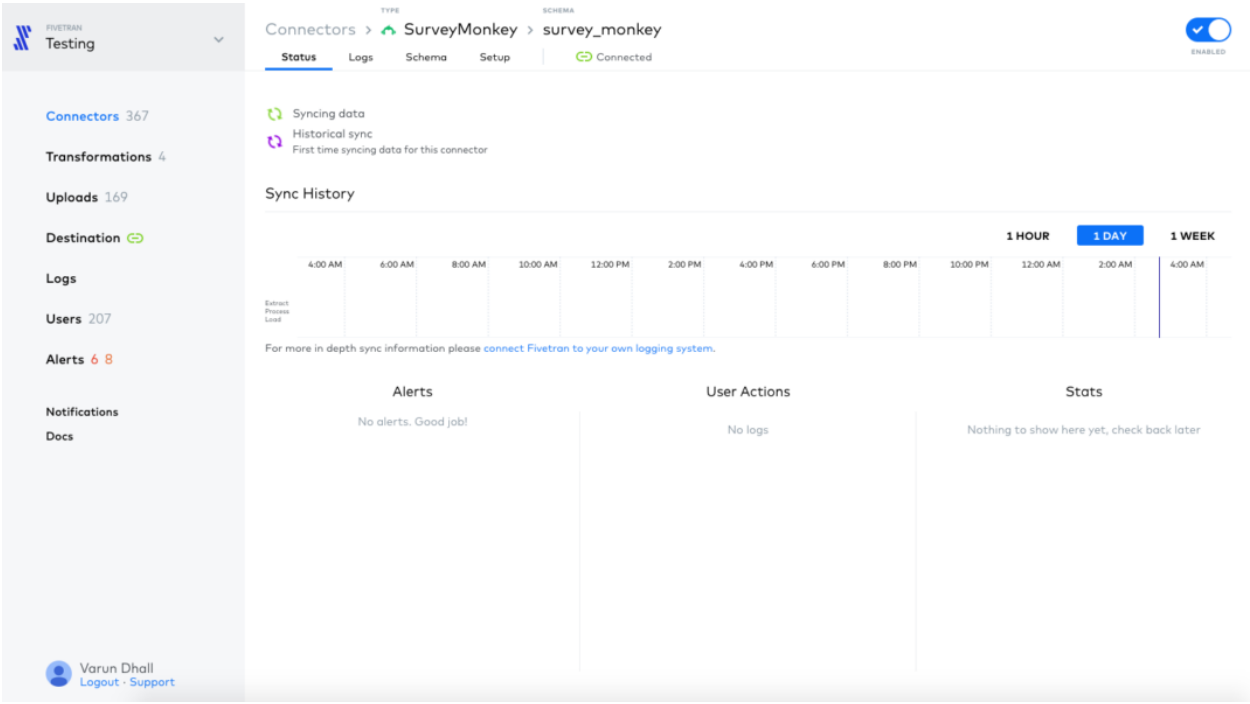
Aim: To study various ETL tools used in data warehouses.
Tools:

1) **Fivetran**

Fivetran is a SaaS data integration company that offers two ETL solutions for businesses and apps. Fivetran can duplicate cloud and on-premises databases, move massive amounts of data, and improve analytics with prebuilt data models, all with 99.9% platform uptime.

- Pros
  I. Permissions for security and administrative access to information that are simple to understand
  II. Syncing data from many databases and cloud apps is simple.
  III. Administrators will benefit from a user-friendly GUI for easy setup and administration.
  IV. In light of the vendor's data processing capabilities, there is a good value for money.
- Cons
  I. The time and effort required for manually resyncing data, as well as the restricted alternatives available
  II. Some supported connections have sporadic response.
  III. The timing of notifications and warnings might be improved.
  IV. There aren't any connectors for some of the most popular data movement apps.
- Features
  I. Soft deletes using log-based replication enable for continuous examination of deleted data by preventing certain columns or tables from replicating to the destination.

II. Use the Fivetran REST API to run central functions for users, groups, and connections.
III. Forward and backward synchronisation steps with priority synchronisations
IV. AWS, Apache, Snowplow, Segment, and Webhooks are all supported by the event tracking framework.



2) IBM InfoSphere Information Server

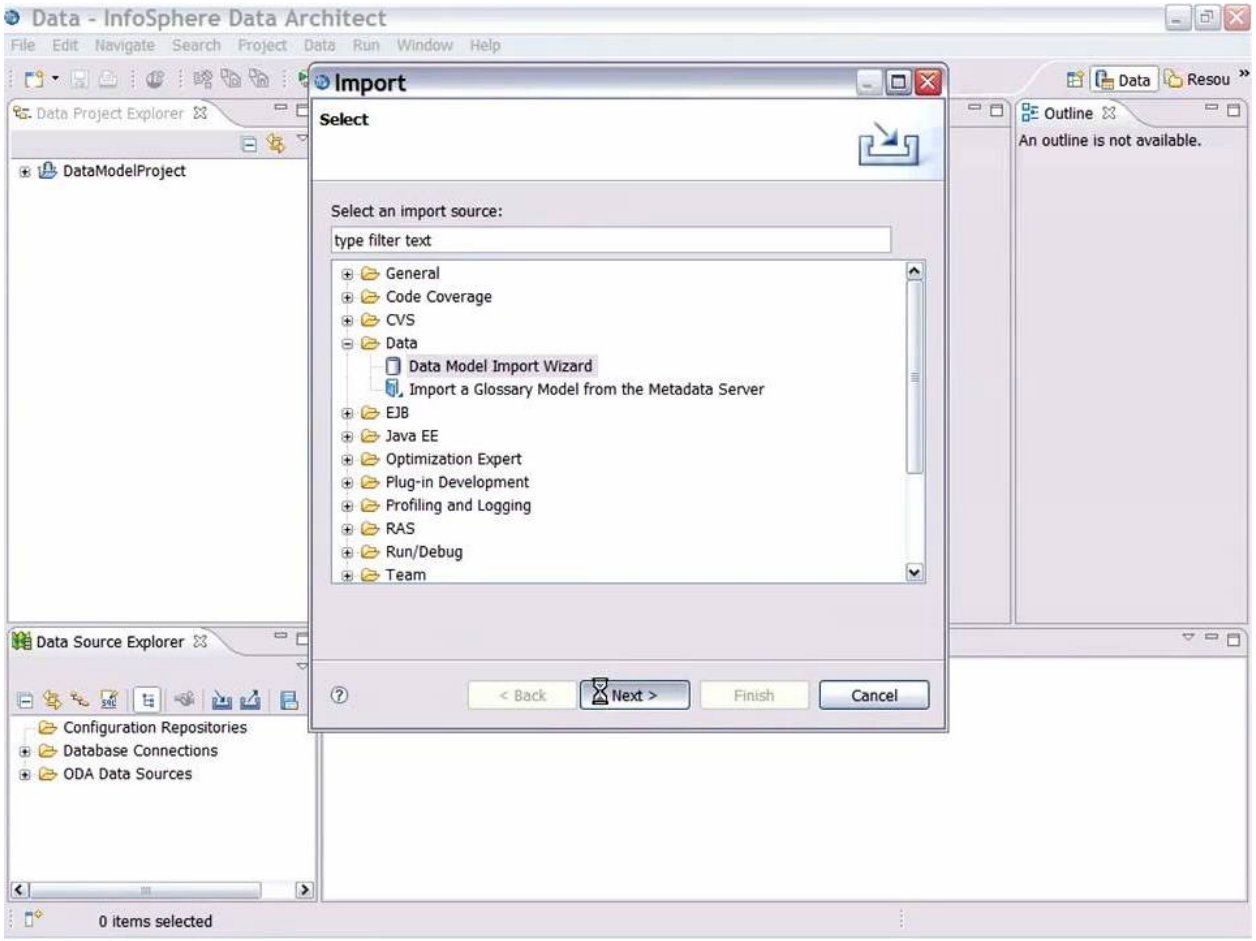
IBM's InfoSphere Information Server is a major data integration platform. The IBM InfoSphere Information Server is an enterprise-ready technology that supports massive parallel processing (MPP). Clients may use an easy online interface to access a variety of capabilities, including multi-cloud data integration, support for unstructured data, and data quality analysis.

- Pros
  I. Existing clients of the vendor's solution stack will find it useful.
  II. Vendor software documentation that can be used and technical support that can be reached
  III. Capabilities for data replication and synchronisation

IV. REST API and flexible, event-driven architecture for client SOA

- Cons
  I. In comparison to other ETL solutions, it's pricey, and it's difficult for small teams to use.
  II. Creating source-to-target maps and assessing multiple workloads is difficult.
  III. There have been a few cases of instability and sporadic responsiveness.
  IV. Due to the difficulty of installation, early configuration management is critical to success.

- Features
  I. Large volumes of data must be extracted, transformed, distributed, and scaled for warehousing.
  II. IBM Watson Knowledge Catalog helps you classify unstructured data sources and automates data quality and governance so you can satisfy compliance requirements.
  III. Maintain data integrity through cleaning, monitoring, and maintaining it.
  IV. Run data profiling and analysis tasks on Hadoop-based big data clusters.



Conclusion:

To create a data warehouse, we need a lot of data from various sources. This data available at different sources may not fit our design. Hence, ETL tools come to rescue from this problem. ETL tools extract data from various sources and transform them into required format making our work easier and faster. They also load the formatted data into data warehouse.