NATIVE DATALAKE; ENABLING A DATA PIPELINE FOR DATA ANALYSIS WITH SMALL TO MEDIUM NATIVE DATASETS

Authors
Ryan Marchand, Todor Petrov, Richard Chapman, Neil Landers,

INTRODUCTION

Small- to medium-sized organizations have a need to analyze their chromatography data across many projects and systems. However, they may find the current ETL (Extract, Transform, Load) solutions to either be infrastructure-intensive, complex or cost-prohibitive. They may not require the same enterprise-level solutions as a large organization. Here, we present Native Datalake: an application leveraging an automated way of centralizing and transforming data for analysis that is more targeted to the needs of small- to medium-sized organizations.

METHODS

Native Datalake is a background application service that leverages the Empower Toolkit in order to set up an ETL layer from a user's Empower™ system to a central database for analysis. Once setup on an Empower system, Native Datalake will regularly synchronize data to the database. This will enable the use of EDA (Empower Data Analytics) web applications to render system utilization, and compliance visualizations by accessing data through Native Datalake. Native Datalake centralizes scientific data from instruments and software systems in a single internal or external database. (Figure 1)

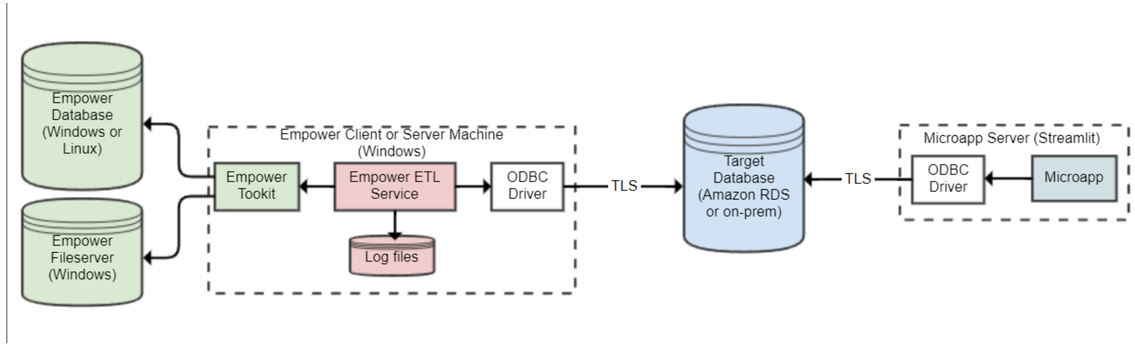

Figure 1. Application Architecture

RESULTS

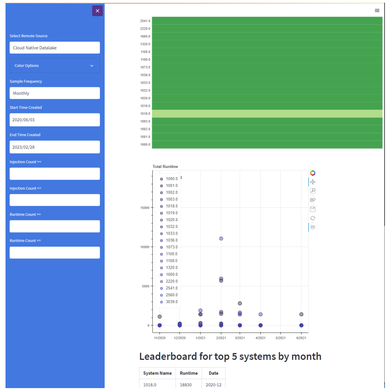

Figure 2. System Utilization

Figure 3. Compliance

CONCLUSION

Leveraging Native Datalake, we created EDA (Empower Data Analytics), an application that provides dashboards of aggregated data from an organization's many systems.

For system utilization(Figure 2), EDA makes database queries to obtain aggregations on system usage and fleet performance. These include:

- **injection counts and total runtime by either project or system rendered onto bar charts.**
- **Data can also be viewed via date hierarchy on a heatmap by year, quarter, month, or day. This enables a user to monitor trends in heavy system usage by date segments such as day of the week. This may justify the need for another system or need for optimizing bottlenecks in day-to-day processing.**

For compliance (Figure 3), EDA identifies potential violations by system, project, and user with SQL queries. These queries include:

- **counts for manual integrations**
- **unprocessed acquisitions**
- **acquisitions that had been processed multiple times**
- **incomplete sign offs and aborted acquisitions.**

By getting a bird's eye view, potential violations are identified across locations. A user can then drill down to the samples, review and confirm the chromatogram and audit trail to identify error or fraud. This ensures business compliance and quality, as well as saving an organization time so they do not have to review every chromatogram individually. This provides value to users like Lab Managers, Lab Analysts and Quality Analysts with optimizing review by exception, quality, safety, efficacy, increasing compliance and audit preparedness.

TO DOWNLOAD A COPY OF THIS POSTER, VISIT WWW.WATERS.COM/POSTERS