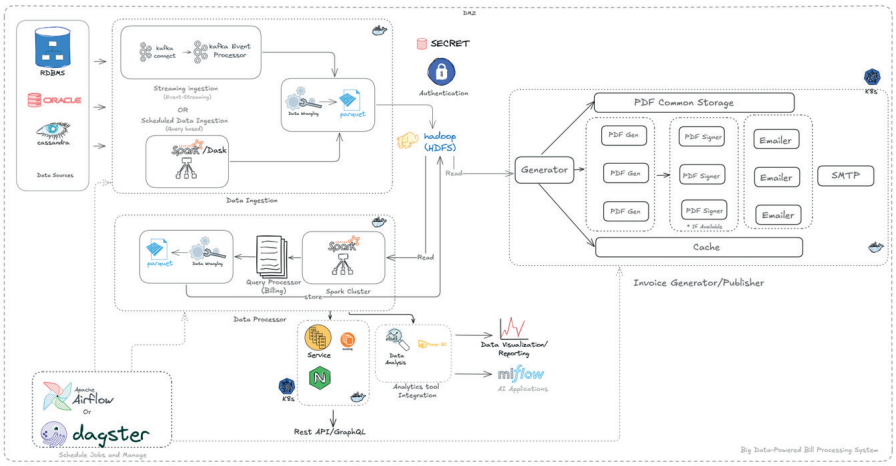
Arun Atri
Head of Offshore Services
aa@eberhardt-partner.com

Andreas Eberhardt
CEO/Managing Director
ae@eberhardt-partner.com

Big Data Powered Bill Processing System

Overview

The product is designed to streamline and modernize invoice generation and analytics using advanced data engineering tools such as Apache Spark, Parquet, Typst, Dagster, and Hadoop. It focuses on handling large-scale invoice generation, PDF creation, and data analysis, offering improvements in performance, storage efficiency, and adaptability compared to traditional SQL-based approaches. Other technologies beyond those mentioned may also be incorporated as needed.



Key Advantages of the Product Over Traditional Solutions

1. Performance Benchmarks:

Task	New Solution *(Using Spark, etc.)*	Traditional SQL-Based Solution
Invoice Generation (2.5M)	30 seconds	390 seconds
PDF Creation (1M PDFs)	30 seconds	3 minutes
Data Analytics Query	25 seconds	400 seconds
Average PDF Size (7 to 15 Pages)	60 KB - 75 KB	2 MB - 4 MB

2. Key Benefits:

- **Storage Optimization:** Significant reduction in PDF size (60 KB - 75 KB compared to 2-4 MB), enabling efficient storage and cost savings.
- **Adaptable PDF Generator:** Automatically adjusts to changes in bill format, eliminating the need for manual updates.
- **Time Efficiency:** Improves processing time by 4-5x compared to traditional invoice generation methods.
- **Database Flexibility:** Seamless integration with relational (SQL-based) and non-relational (NoSQL) databases.

3. Enhanced Capabilities:

- Conversion of stored procedures to Apache Spark queries for enhanced performance.
- Built-in API for seamless billing data distribution across systems and services.
- Easy integration with Power BI for advanced invoice and revenue analytics.
- Capability to merge customer bills from multiple billing systems.
- Supports machine learning and data analytics processes such as churn prediction and fraud detection.
- **User Convenience with Dagster:** The Dagster workflow orchestration enables one-click processing for streamlined operations, ensuring the best user experience.

4. Operational Cost Reduction:

- Reduced processing and storage requirements significantly lower operational costs.
- Optimized resource utilization through distributed computing and efficient data storage formats like Parquet.

Technical Stack

- **Apache Spark:** High-performance distributed data processing for invoice generation and analytics.
- **Parquet:** Efficient columnar storage format for optimized data storage and retrieval.
- **Typst:** Modern PDF generation tool for creating adaptable and compact invoice PDFs.
- **Dagster:** Workflow orchestration tool for managing and scheduling data processing pipelines.
- **Hadoop:** Distributed file system for handling large-scale data storage and retrieval.

Additional Advantages

- **Scalability:** Designed to handle millions of invoices with ease, ensuring scalability for growing business needs.
- **Real-Time Insights:** Enables near real-time analytics for informed decision-making.
- **Security:** Ensures secure processing and storage of sensitive billing data.
- **Cross-Platform Compatibility:** Connects with diverse systems and technologies to offer a unified billing experience.

Future Scope

- Integration with AI models for predictive analytics (e.g., demand forecasting).
- Real-time error detection and self-healing pipelines for continuous operation.
- Advanced visualization dashboards for dynamic reporting and trend analysis.

This solution offers a comprehensive, efficient, and modern approach to invoice generation, delivering unparalleled advantages over traditional methods.