Google Cloud Platform

A Cloud Architect’s Handbook:

How to Build an End-to-End Cloud Data Integration Solution Using Talend and Google Cloud Platform

talend

INTRODUCTION

Talend is a leading cloud and big data software provider for data-driven companies, offering a single platform for data integration, data management, and application integration & API delivery use cases that delivers agile analytics across public, private, hybrid, multi- clouds as well as on-premises environments.
Talend is a leader in the 2017 Gartner Magic Quadrant for Data Integration Tools and the Forrester Wave: Big Data Fabric Q4 2016, and 2018; its solutions support over 2,500 global enterprise customers across multiple verticals.

As a strategic partner with Google, Talend provides native connectivity to numerous Google Cloud Platform services such as Google Cloud Storage, Google BigQuery, and Google Dataproc. This enables businesses to leverage the joint solution to create cloud data lakes, build cloud data warehouses, perform data governance, modernize data management and data integration platforms — all to ultimately quickly gain trusted insights with speed and agility at a predictable price.

By combining the power of Talend and the Google Cloud Platform, many customers have successfully built an end-to-end cloud analytics solution for their integration projects. This paper describes the customer business use cases and reference architectures used in building material supplier, online advertising industries and manufacturing and retail industries.

TABLE OF CONTENTS

Use Case 1: A Building Materials Supplier Increasing Online Sales Conversion Rate with Higher Quality Data

The U.K.’s largest global building materials supplier was experiencing enormous success with their online business. But they urgently needed to get better data to bring new products online faster, more accurately demonstrate the breadth of their offerings, and come up with a better marketing strategy to drive more online sales. However, their current data, distinctly siloed, is neither maintained or validated in a consistent way for accurate decision making.

Prioritizing flexibility, ease-of-use, data quality capabilities and pricing in the vendor selection criteria, the company selected Talend Data Fabric — a single, unified data integration and management platform across cloud and on-premises environments, to tackle their data challenges. Talend, together with the Google Cloud Platform, were able to successfully consolidate the company’s data from thousands of branches, hundreds of outlets and half-million product lines into a data lake and validate it in a consistent manner. As a result, the company saw a 30% increase in sales conversion rate only one month after the solutions were implemented.

Use Case 1: A Building Materials Supplier Increasing Online Sales Conversion Rate with Higher Quality Data

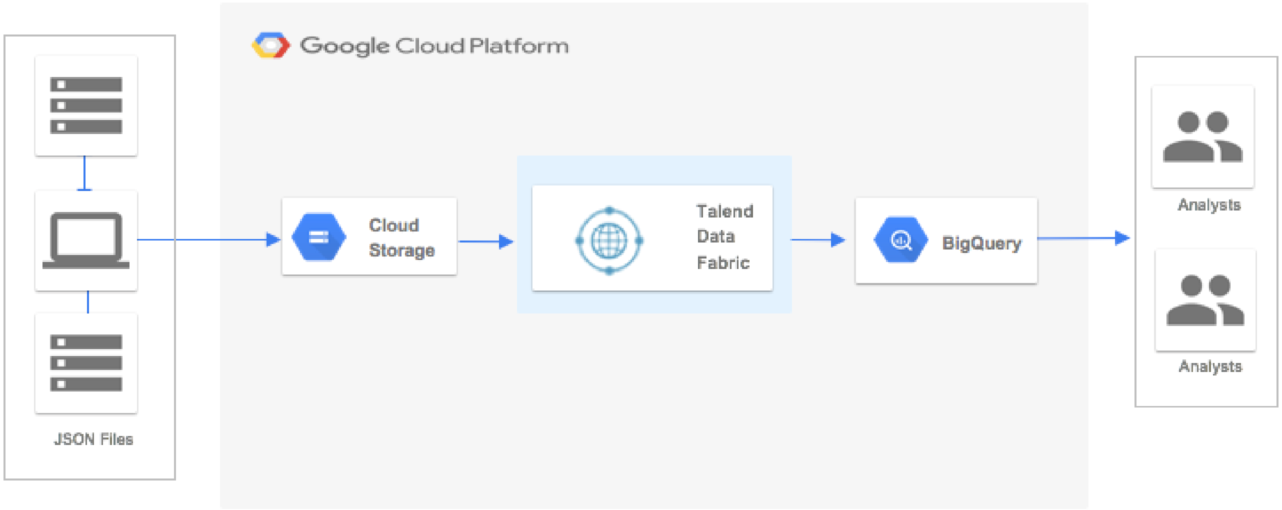

Figure 1: Cloud Data Lake Architecture using Talend Data Fabric and the Google Cloud Platform

Data Ingestion

The data was collected from the company’s hundreds of Product Information Management (PIM) systems, internal transactional systems, and external partner and customer systems across the country. The data was mainly in JSON flat files containing information on invoices, products, and other business metrics. The company used Talend to merge millions of records of raw data from different PIM systems into a single, completed copy and upload them to Google Cloud Storage, a secure RESTful online file storage system for storing and accessing data on Google's infrastructure.

Data Transformation and Data Governance on the Google Cloud Data Lake

To improve data quality, the company integrated Talend Data Services into their PMI tools to provide standardization, validation, and normalization of data. In essence, Talend Data Services offers built-in data quality capabilities that provide a data quality firewall—so any data entering the company’s systems first run rules to check for duplicates, confirms that check digits are valid for barcodes, standardizes data, and maintains a consistent master list of values.

Meanwhile, the Talend Data Integration component of Talend Data Fabric performs data transformation. Once the quality of the data in the files meets established standards, it is sent to Google BigQuery for analysis and reporting.

The company uses Talend Data Stewardship to correct data before it's routed to other systems. This further ensures that all data that are sent to the target data apps are accurate and reliable for analysis. They then use Talend Data Preparation to provide self-service for ETL processes by business users, enabling them to define mapping or standardization rules themselves rather than relying on a developer to do so.

Continuous Integration and Master Data Management

The company also uses Jenkins Continuous Integration to publish jobs in Talend Cloud and Talend Metadata Manager and Talend Master Data Management at the enterprise level for data search and discovery.

Products in This Architecture

- Talend Data Fabric
- Talend Big Data Platform
- Talend Data Preparation
- Talend Data Stewardship
- Talend Data Services Platform
- Talend Master Data Management
- Google Cloud Storage
- Google BigQuery

Use Case 2: A Media Publisher Improving Online Advertising Revenue Streams by Building a Single Customer View in the Cloud

A leading online automotive media publisher is a platform that gathers auto listing data from multiple sources and creates award-winning content and services to help their automotive clients grow. The company uses Google Analytics to capture and analyze clickstream data. To enhance its advertising revenue streams, it is critical to accurately understand customers' browsing behaviors on their website, customize their advertising content, point the customers to best-matched car dealers, and get insights on why they make a purchase from one dealer versus another.

To get those insights, the company needs to collect clickstream data from the website where it posted classified ads for autos, as well as aggregate data from other sources, including car dealerships, ad agencies, and other internal on-premises databases for real-time and ad hoc analytics. However, their current ETL tools limited their ability to handle their Big Data volume. They ultimately selected Talend Big Data Platform and the Google Cloud Platform to spin up a BI (business intelligence) lake to solve their integration problems.

The lake was set up within 4 weeks and provided the company with a single view of all relevant data for deeper analysis.

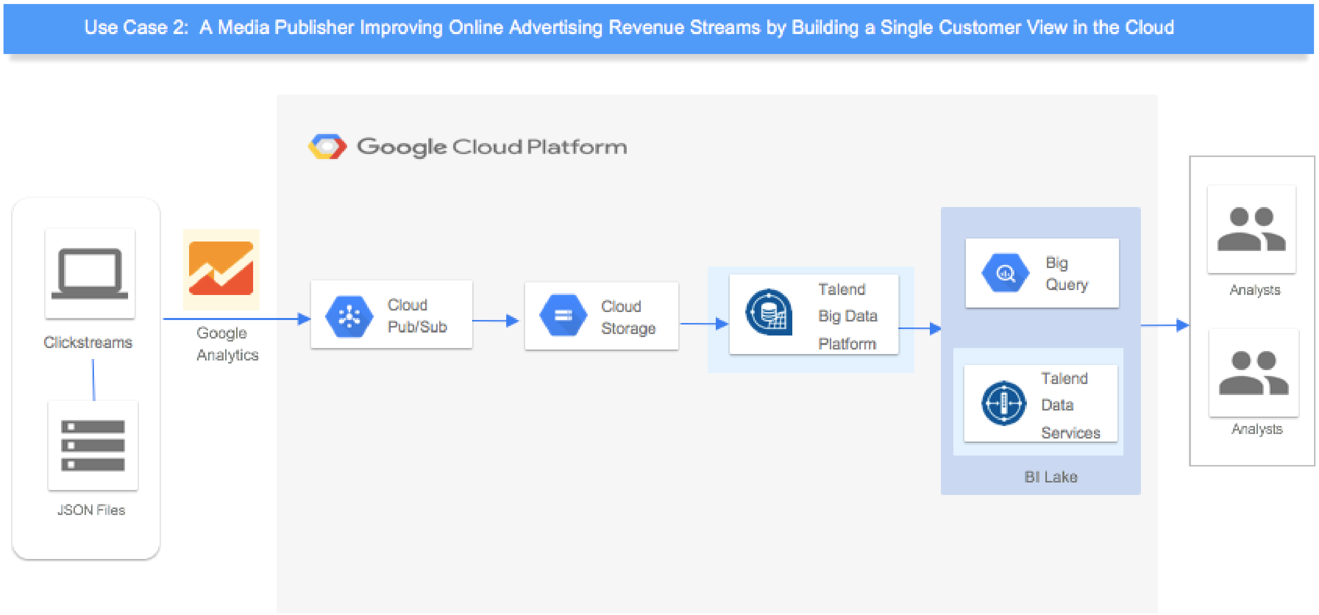

Figure 2: Real-Time Processing Data Workflow using Talend and Google Cloud Platform

Data Ingestion

Much of the customer’s data is XML payloads and JSON payloads from Google Analytics data. They first use Google Cloud Pub/Sub — a fully-managed real-time messaging service that allows you to send and receive messages between independent applications — to upload real-time data from clickstreams and external applications from partners directly into Google Cloud Storage and parse it there.

Data Transformation

Data is sent from Google Cloud Storage to the BI lake — which consists of Google BigQuery and Talend Data Services —passing first through the Talend Big Data Platform job server, which does the transformation, enrichment, and cleansing. Talend Big Data Platform also joins datasets and turns complex data structures into simple, rule-based data structures; for example, it can extract complex JSON payload data into a flat row structure to be stored in Google Cloud BigQuery. BigQuery is an enterprise cloud data warehouse that works in conjunction with Google Cloud Storage and enables fast SQL queries for analytics with built-in machine learning.

BI Lake for Data Analysis

The data residing in the BI lake for use by BigQuery is partitioned into different tables, and users then run queries across tables. Talend Data Services enables business users to access the API, thus making the process more event-driven. Data is then passed to Tableau for analysis.

Products in This Architecture

- Talend Big Data Platform
- Talend Data Services Platform
- Google Cloud BigQuery
- Google Cloud Storage
- Google Cloud Pub/Sub

Use case 3: A haircare and skincare company builds a BI lake to streamline operations

A fast-growing hair care and skincare enterprise, recognized for its innovative use of high quality and authentic ingredients to serve the unmet needs of consumers of color, recently got acquired by a large global brand. Due to new market opportunities through the acquisition, there was an urgent need for the company to become data-driven and agile and to scale faster.

To achieve that, they needed a new IT solution that could create a new workflow process and a cost-efficient data analytics architecture that could allow them to work with real-time data. This will enable them to address the challenges that most manufacturing and retail vendors face: projecting daily sales volumes, monitoring sales across different stores and geographies, planning for shipment and logistics, managing human resource and shifts, and managing vendor costs.

The Solution Selection Criteria

The customer had been traditionally operating on an ERP solution based on an SQL server and XLS sheets for most of their business data needs. After consulting a technology services firm, the client decided to create a data lake that could meet the following requirements:

1. Scalable
2. Open source first, no vendor lock-in
3. Low cost
4. Flexibility to migrate data between data lake technologies
5. Secure and protected
6. SQL on Hadoop | EDW on parallel database
7. Ability to ingest structured, quasi-structured and file data from multiple sources
8. Flexible and multi-source pipeline management
9. Support for NoSQL databases like MongoDB
10. 24 x 7 x 365 Operational Support

The data integration and management platform needed to be able to provide the following features:

1. Open source and comes with enterprise support
2. Multi-user & version control capable
3. Data quality and governance capable
4. Master Data Management capability
5. Works with third-party vendor sources for data ingestion6.
6. Ability to process batch, stream, and data in-memory

Cloud Data Lake Architecture

After evaluating several vendor solutions, the company decided to go with Talend Big Data Platform, as the Talend platform provided all the desired features, including open source capability. The solution itself was also capable of being virtualized and could work with disparate sources of data. Talend's ability to work with multiple vendors allowed the solution to be future-proof and allowed the client to plan a platform road map.

To keep the cost low, the company planned to run processes requiring a large number of computing resources and prescriptive analytics on Google Cloud DataProc and Google Cloud BigQuery. They designed and deployed a completely virtualized solution on the Google Cloud Platform. This architecture created actionable business data every hour and on demand.

Use Case 3: A Haircare and Skincare Company Builds a Business Intelligence (BI) Lake to Streamline Operations

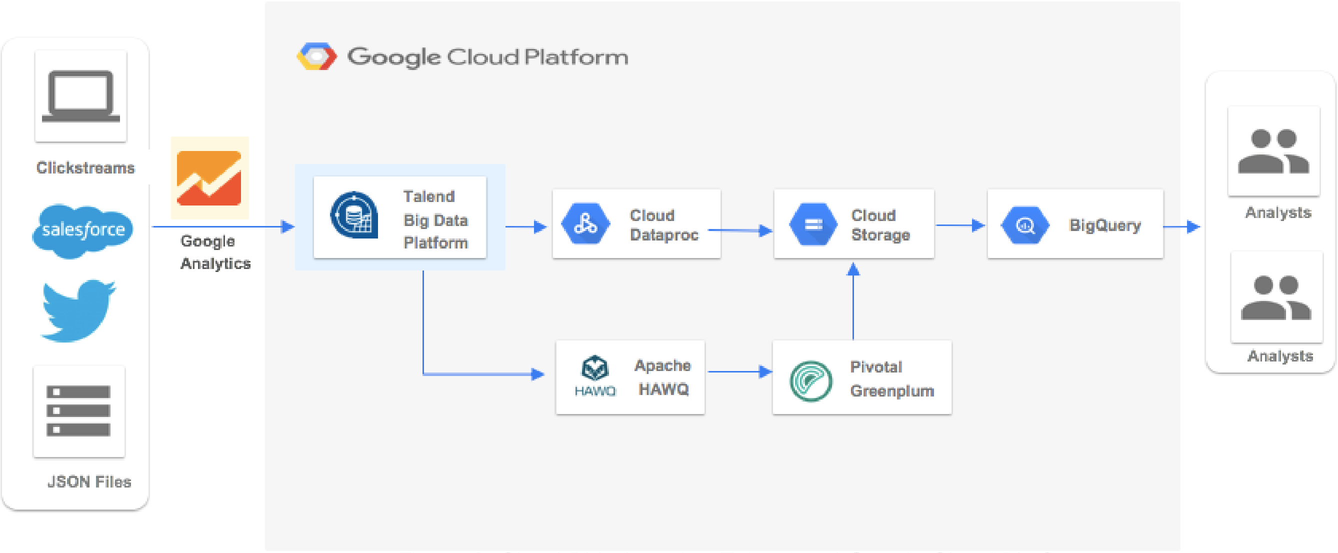

Figure 3: Cloud BI Lake using Talend and Google Cloud Platform

The architecture starts with ingestion. Data is first collected from the client's ERP systems (SQL), Salesforce systems, and social media sites, and then sent to Talend Big Data Platform and Google Cloud DataProc for processing. The batch processes that needed to run on a 24 x 7 basis were planned on Hortonworks Hadoop and Pivotal Greenplum Database (MPP Database). This architecture was further layered with Google's virtual network and VPN solutions. This has allowed for a fully protected data architecture. The processed data is then sent to Google BigQuery for analytics and BI reporting.

The Business and IT Outcome

On a broader scale, the following outcomes were achieved by the customer:

1. Stream and batch process 30GB+ data/hr. in the pipeline
2. 100+ parallel jobs to plan and forecast manufacturing and sales
3. Hourly logistic plans
4. Daily scheduled reports on end-of-business sales

5. Automated KPI matrix from verifiable sources
6. Master Data Management
7. Data Security and Governance
8. 24 x 7 x 365 support for Data platform and Business Intelligence

This low cost, highly efficient footprint has allowed our customer to continue towards their growth path and be an innovative technology leader in the manufacturing and retail space.

Products in This Architecture

- Talend Big Data Platform
- Talend Master Data Management
- Google Cloud Storage
- Google BigQuery
- Google DataProc

Talend: A Cloud Data Integration Leader

Talend (Nasdaq: TLND), a leader in cloud integration solutions, liberates data from legacy infrastructure and puts more of the right data to work for your business, faster. Talend delivers a single platform for data integration across public, private, and hybrid cloud, as well as on-premises environments, and enables greater collaboration between IT and business teams. Combined with an open, native, and extensible architecture for rapidly embracing market innovations, Talend allows you to cost-effectively meet the demands of ever-increasing data volumes, users, and use cases.

Over 1,500 global enterprise customers have chosen Talend to put their data to work including GE, HP Inc., and Domino's. Talend has been recognized as a leader in its field by leading analyst firms and industry publications including Forbes, InfoWorld, and SD Times. For more information about our work with the Google Cloud Platform, and to learn more about how we could put more data to work for your business, contact us today.

Contact us: www.talend.com/contact
Facebook: www.facebook/talend

Twitter: @talend