A Cloud Architect’s Handbook:

How to Modernize Your Cloud Platform for Big Data Analytics with Talend and Microsoft Azure

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 use cases that delivers agile analytics across public, private, and hybrid clouds as well as on-premises environments.

In partnership with Microsoft, Talend provides fast development of Big Data ETL processing, cloud data lakes, cloud data warehousing, and real-time analytics projects on the Microsoft Azure Cloud Platform. This empowers companies to solve modern integration and analytics challenges by connecting business-critical data and applications from on-premises systems, cloud, social, and mobile apps in real-time at a predictable price.

By combining the power of Talend and Microsoft Azure, many organizations have successfully modernized their cloud platforms for big data analytics. This white paper details use cases in the energy, food, beverage & brewing, and logistics industries, as well as the IT architectures that were used in the solutions.

TABLE OF CONTENTS

USE CASE 1: Maximizing Customer Engagement to Keep a Liquid Petroleum Gas Supplier ahead of the Competition

Maintaining a high level of customer engagement is critical to keeping the competition at bay for any company, yet for a leading British liquid petroleum gas supplier, it requires a Herculean effort. They must keep customer engagement high across a number of criteria ranging from product quality to pricing to supply and operations to a compelling branding and positioning strategy. One way to ensure great customer engagement is to find the right customer segment and target them with the right messaging at the right time through the right channel. The challenge, however, lies in getting relevant, accurate, and in-depth data of individual customers.

Using Talend Big Data Platform to build a cloud data lake on the Microsoft Azure Cloud Platform, this company was able to integrate and cleanse data from multiple sources and deliver real-time insights. With a clear view of each customer segment's profitability, they could target their customers with customized offers at the right time to maximize engagement.

The cloud data lake architecture consists of Talend Big Data Platform, Microsoft Azure Data Lake Store (ADL Store), Azure HD Insight, and Azure SQL Data Warehouse. The architecture allows the company to move massive amounts of data from several on-premises apps into a central cloud repository on Azure for real-time analytics.

Data Ingestion

The gas supplier's files are on the JD Edwards EnterpriseOne (E1) ERP system in several locations. The system lacks data validation capabilities and contains records with incorrect data entered by users. The company also uses TIBCO Enterprise Service Bus (ESB) to extract, add data, and detect changes in the data within the E1 ERP system. An ESB is an architecture with a set of rules and principles for integrating numerous applications together over a bus-like infrastructure. Talend Big Data Platform downloads the files from E1 and then moves them to Microsoft Azure Data Lake Store (ADL Store), an enterprise-wide hyper-scale repository optimized for Apache Spark and Hadoop Analytics engines. It offers a single storage for file system and object data that is fully integrated with Azure Blob Storage.

Data Transformation

Talend Big Data Platform quickly integrates, cleanses, and profiles the ingested data stored on ADL Store, while the customer adds requirements for data governance, business rules, and compliance rules. The data is then sent to Azure HDInsight, a service that enables clusters of managed Hadoop instances and is commonly used for easy, fast, and cost-effective big data processing. The process of ingesting data into Azure using Talend with this architecture is 50% faster than from their existing ETL architecture.

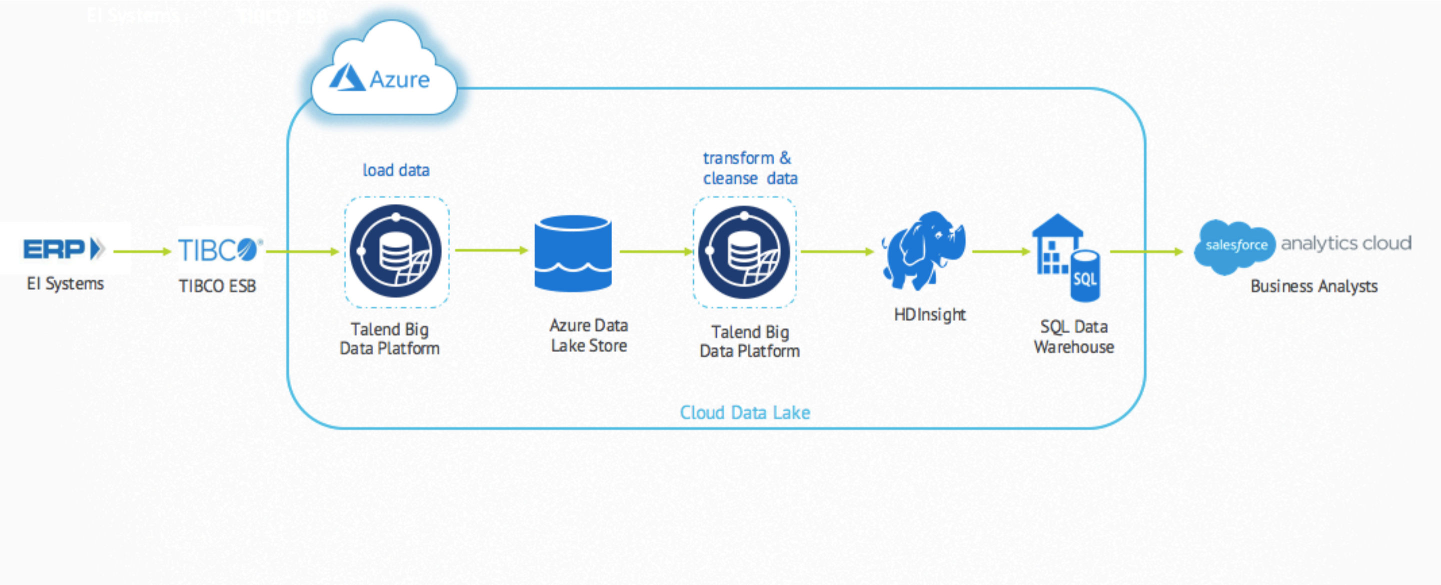

Figure 1. Azure Cloud Data Lake for Maximizing Customer Engagement

After the customer applies rules, Talend Data Quality checks data in fields such as address, post codes, names, phone numbers, and other reference fields. It then verifies the customer business sectors that are encoded in E1: agriculture, domestic and manufacturing, for example. When needed, the data team uses Talend Data Stewardship to manually correct data before it's routed to other systems. This further ensures that all data that is sent to the target data apps are accurate and reliable for analysis. Customers can add more data quality rules to systems residing in Azure as they scale, and get a single view of their customer data using Talend Master Data Management (MDM).

Data Warehouse for Analytics

After the data is cleansed and transformed, Talend Big Data Platform then bulk loads the validated files to Azure SQL Data Warehouse, a powerful, fully managed, petabyte-scale data warehouse. Salesforce Analytics then pulls the cleansed and complete data for reporting and analysis.

Products in This Architecture

- Talend Big Data Platform
- Talend Data Stewardship
- Talend Data Quality
- Talend Data Mapper
- Talend Master Data Management (MDM)
- Microsoft Azure Data Lake Store (ADL Store)
- Microsoft Azure HDInsight
- Microsoft SQL Data Warehouse

USE CASE 2: Enabling GDPR Compliance and Social Media Analytics to Improve Marketing Campaigns for a Food, Beverage & Brewing Company

Balancing visibility into customer data in order to design effective marketing campaigns while complying with data regulations is not an easy task for the highly-regulated liquor industry, as wine, beer, and spirits companies are not allowed to collect customer or retail store data first-hand with surveys.

The CTO of a century-old large European food, beverage and brewing company with 500 brands was able to achieve this balance, however, with a GDPR-compliant solution that delivers insights on how customers and prospects talk about their products and services on social media platforms in real-time.

Using Talend Big Data Platform and Microsoft Azure to build an enterprise cloud data lake, the company was able to analyze various social media data from 450 topics with a daily sample set of up to 80GB and transform over 50 thousand rows of customer data in a time span of 90 days.

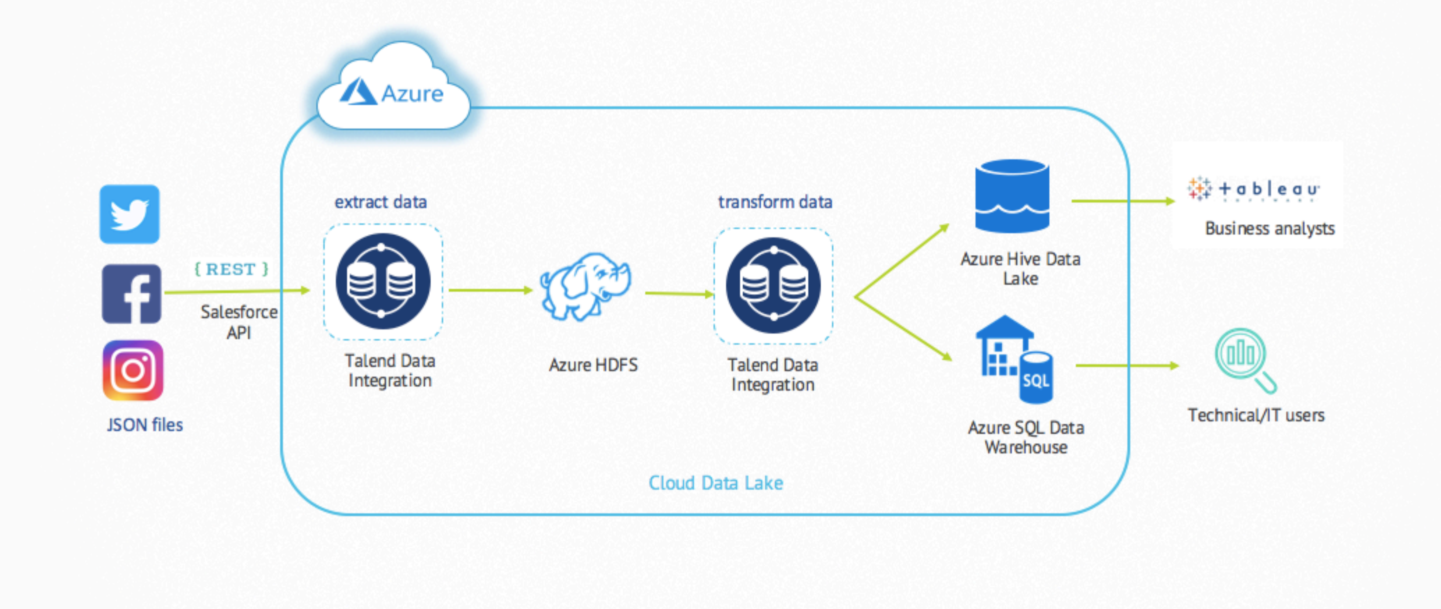

Figure 2: Azure Cloud Data Lake for Social Media Analytics

The main goal of the architecture is to extract all relevant posts from social media sites such as Instagram, Facebook, and Twitter and load those data into the Azure data lake, built with Azure HDFS and Azure Hive Database. From there, the organizations use the transformed data for analytics.

Data Ingestion

The company defined 450 topics to analyze and set a rule to extract only data from the past 90 days, as older data doesn't provide much insight into current trends and customer sentiments. For the extracted data, Talend Big Data Platform first ingests the raw data in JSON format through the Salesforce REST API using the appropriate authentication parameters. The data contain User ID, comments, sentiments, and the numbers of likes and repostings. Talend Big Data Platform removes the content that includes symbols and characters, then converts the data into flat files and filters out only the information needed. In the data ingestion phase, the customer applies business rules. For example, if the author value is null, they define it as - 99 (minus 99). Other transformation rules include classification types and engagement types, which are used to filter data based on directories. Talend Big Data Platform then bulk loads the transformed files into Azure Hadoop File System (HDFS).

Data Transformation, Analytics, and Auditing

In Azure HDFS, the Talend job server transforms the data and sends it to 2 locations: Azure Hive Data Lake that is built on Hive database models and Azure SQL Data Warehouse. A team of data scientists and business analysts then pull the data from Azure Hive Data Lake to build marketing campaigns. For another project, the company integrated the cleansed social media data with other data sources to serve other departments: finance for demand forecasting, farming to predict irrigation needed to produce the best crops, operations for supply chain optimization and many more. In parallel, the data sent to Azure SQL data lake is used for internal IT and auditing needs.

Data Governance and Compliance

Data compliance and governance were a top concern for the retailer. Failure to comply with the EU GDPR can result in a large financial penalty. Therefore, they leveraged Talend to build a strong data governance program to ensure compliance, as Talend maps relevant data elements across datasets using metadata, designs and operationalizes data controls all along the data pipelines, and tracks and manages data with audit trails and data lineage.

Products in This Architecture

- Talend Data Integration Platform
- Talend Data Stewardship
- Microsoft Azure HDFS
- Microsoft Azure Hive Data Lake
- Microsoft SQL Data Warehouse

USE CASE 3: Delivering Real-Time Package Tracking Services by Building a Cloud Data Warehouse

To maintain a premium level of package tracking and delivery service, a leading logistics solution provider needed to consolidate, process and accurately analyze raw data from scanning, transportation, and last mile delivery from a wide range of in-network applications and databases.

They selected Talend for its open source and hybrid nature, its developer-friendly UI, and simple pricing. By deploying Talend Real-Time Big Data Platform the Microsoft Azure cloud environment, they were able to rearchitect a legacy infrastructure and build a modern cloud data warehouse that allows them to provide cutting-edge services, and shrink package tracking information delays from 6 hours to less than 15 minutes.

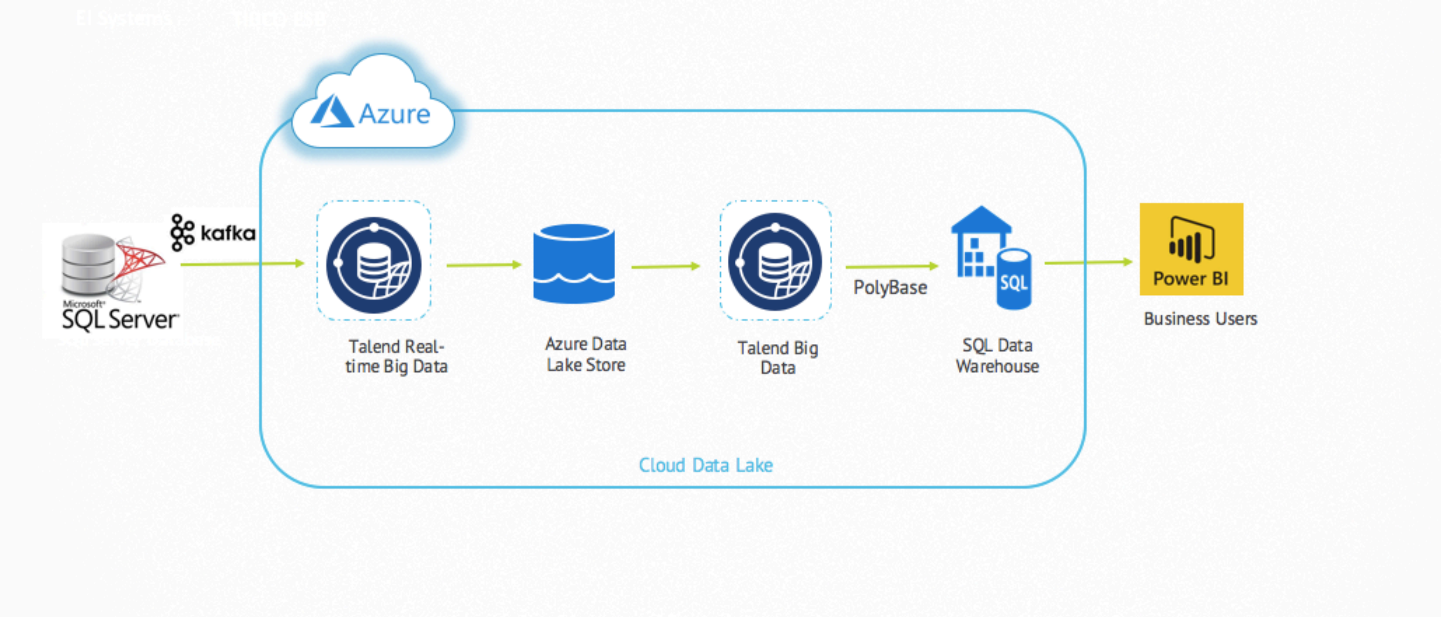

Figure 3: Cloud Data Warehouse to Deliver Tracking Information Real-Time

Data Streaming with Kafka

Hundreds of millions of rows of data are stored across 14 on-premises SQL server databases in the format of text files and parquets. Using a custom Java connector for Kafka, the data is loaded into Kafka transformation engine. Kafka is a streaming-processing software used for building real-time data pipelines and streaming apps and is open source, horizontally scalable, and fault-tolerant. Kafka allows multiple queries for package low latency tracking. All of the Kafka jobs are managed by Talend Administration Center (TAC), a web-based application that centralizes the management and administration of the users' roles, access rights, and job scheduling and monitoring. Then the .NET transformation processors grab all events and send them into Azure cloud SQL Database, a managed cloud database provided by Azure. The customer also used PolyBase to query data in the SQL server.

Data Transformation

Talend Real-Time Big Data Platform then performs the partitioning, transformation, and cleansing, and bulk loads the data from the SQL database to Azure SQL Data Warehouse, a powerful, fully managed, petabyte-scale cloud data warehouse built for reporting and analytics.

Target: Power BI for Analytics

Once the partitioning, transformation and real-time processing are complete, the data is sent to Power BI, their BI tool, and Talend Data Preparation cloud is used to extract data for their day-to-day needs. In the end, this architecture built with Talend and Microsoft Azure is able to deliver the near real-time tracking information that the whole cloud data warehouse project promises.

Products in This Architecture

- Talend Real-Time Big Data Platform

- Talend Data Preparation
- Microsoft Azure SQL Data Warehouse

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 Microsoft Azure, 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