HOW LEADING ENTERPRISES ACHIEVE BUSINESS TRANSFORMATION WITH TALEND & AWS

Cloud Architect's Handbook

talend

INTRODUCTION

Talend is a leading data integration and data management solution provider for data-driven companies. As an Advanced Technology Partner in the Amazon Web Services Partner Network, Talend provides fast development of Big Data, real-time analytics and ETL projects on Amazon Web Services, empowering companies to solve modern integration challenges by connecting business-critical data and applications from on-premises systems, cloud applications, web, social, and mobile apps in days at a predictable price.

By combining the power of Talend and AWS, many customers were able to successfully transform their businesses. This paper describes use cases in the pharmaceutical and food and beverage industries, as well as the IT architectures that were used in the solutions.

TABLE OF CONTENTS

USE CASE 1: TRANSFORMING OPERATIONAL REPORTING SYSTEMS IN A GLOBAL PHARMACEUTICAL COMPANY

To set a long-term growth strategy, a leading pharmaceutical company urgently needed to consolidate multiple operational systems globally and to enable timely delivery of accurate KPI data to provide needed insights.

They partnered with Talend for its fast, unified, and scalable data integration services. By deploying Talend Data Integration and Talend Application Integration in the Amazon Web Services (AWS) cloud environments, they designed and built a modern, cloud- based operational reporting infrastructure that delivers better data insights more quickly at minimal cost.

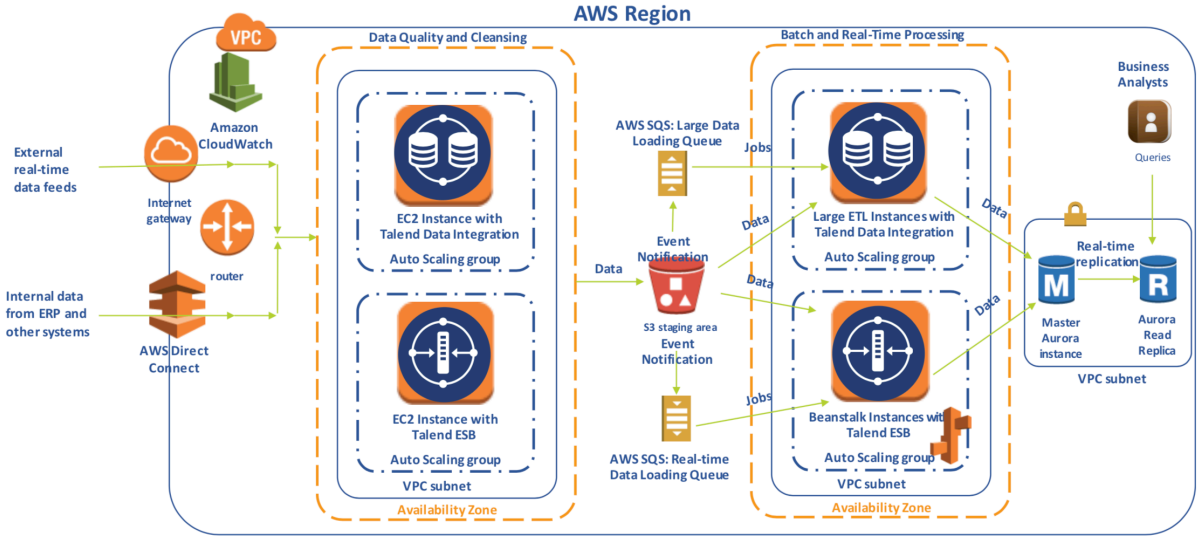

Figure 1: Operational Reporting Infrastructure with Talend and AWS

The entire operational reporting infrastructure is architected using AWS Virtual Private Cloud (VPC) environments. It comprises of separate areas for data ingestion, data quality, staging, ETL, ESB, and finally pushing the data into a target relational database.

Data Ingestion

Internal data from ERP systems such as SAP and other systems is loaded into the customer's data lake using a secure VPN. The VPN effectively makes the AWS VPC an extension of the corporate network and, as such, applications and data transfers are secure and cost effective as reduction in internet traffic reduces costs. AWS VPC is an on-demand configurable pool of shared computing resources allocated within the public AWS cloud environment, providing a higher level of security and control between the different organizations. External real-time data feeds go through the Internet Gateway such as SFTP or via a push from the client's middleware service. Talend ESB is used to integrate services and APIs without any coding. It also simplifies complex mapping challenges and delivers enterprise security.

Data Profiling and Cleansing with Talend

Once all of the data is ingested into the data lake on AWS, event notifications will take place that put messages on a queue. These queues allow AWS EC2 instances to process the data package. The instances are part of AWS Elastic Beanstalk, a free cloud deployment and provisioning service, which is used to automatically and elastically process any new real-time data feeds. Talend Data Integration is used to quickly integrate, cleanse and profile the ingested data. It also synchronizes the metadata and delivers self-service data preparation. The Beanstalk configurations allow instances to process different tasks. One instance type deals with the real-time data feeds, the other with the large scale batch data. Each instance is one auto-scaling group. As auto-scaling groups can detect unhealthy instances and adjust capacity as needed, adding it to the infrastructure is one way of building better fault tolerance, better availability, and better cost management.

Batch and Real-Time Processing

The cleansed data is then moved to the data lake on AWS. From there, the data is fed with two sets of AWS Simple Queue Services (Amazon SQS), a messaging queue service that handles messages or workflows between other components in a system. Different queues handle different data sets. As such, this allows for performance in loading data and also means that batch jobs are separated from real-time transactional loads. AWS Elastic Beanstalk is again used to automatically and elastically process any new real-time data feeds. Whenever new data feeds come in, a Talend ESB Runtime will be spun up to process it. Similarly, whenever batch data comes in, Talend Data Integration will process all of it and after processing, both sets of data are sent to the Operational Data Warehouse (ODS), which is a database designed to integrate data from multiple sources for additional operations on the data.

Target: Relational Database

Once batch and real-time processing is completed, the data is sent to the data target area (ODS). In this case, the customer uses multiple AWS Aurora instances. The first instance is the master instance and the others are replication instances used for continuously reading replicated data. The second instance is where business users

would use their own BI tools or Talend Data Preparation, which is currently being tested, to extract the data for their day-to-day needs.

PRODUCTS IN THIS ARCHITECTURE

- Talend Data Integration
- Talend Data Services Platform (ESB)
- AWS Aurora
- AWS Elastic Beanstalk
- AWS S3
- AWS CloudWatch
- AWS Virtual Private Cloud (VPC)
- AWS Simple Queue Service (SQS)
- AWS Auto Scaling Group

USE CASE 2: ANALYZING PUBLIC DATASETS FROM CLINICAL TRIAL STUDIES TO IMPROVE MEDICAL TREATMENTS

Bringing a new drug to market can be a large undertaking for any pharmaceutical company, ranging from the expense associated with R&D to the approvals required from the FDA. When it comes to treating serious conditions such as tumors and cancers, a Herculean effort is needed to ensure that the drug will become a success. One of the ways to accelerate time-to-market for a new drug is to make use of publicly available datasets consisting of detailed clinical studies. The challenge however, lies in the complex and hierarchical nature of such

datasets, which often tend to be in the form of very large XML files.

Using Talend Big Data Platform and AWS, a pharmaceutical company was able to build a quick proof-of-concept to streamline and transform over 52,000 XML files of clinical study data.

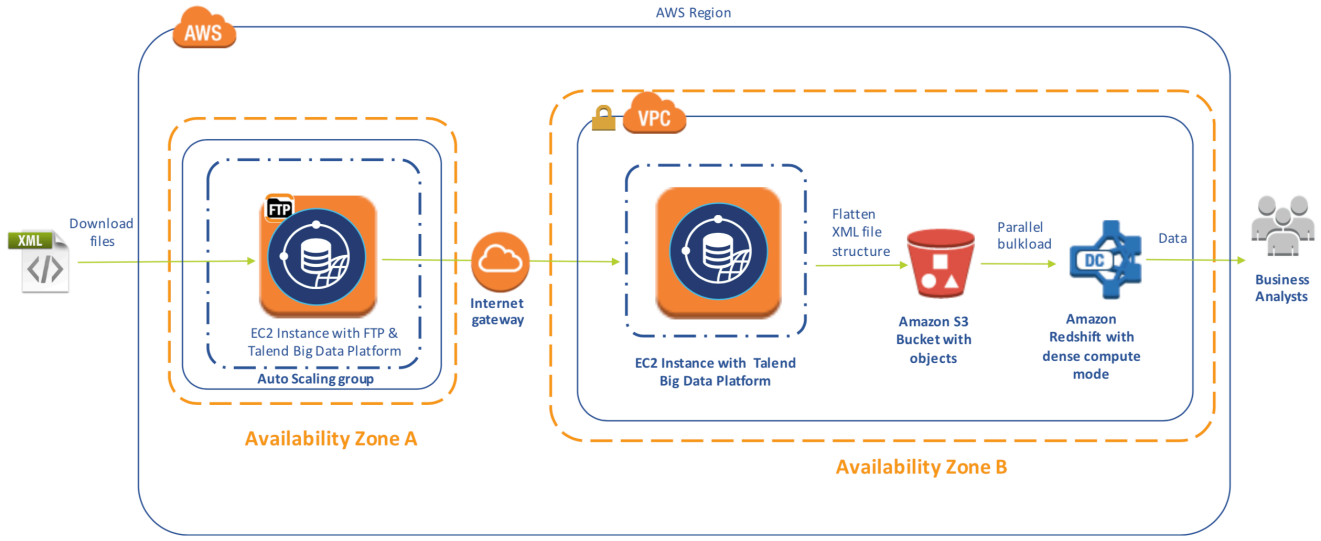

Figure 2: Optimizing Medical Treatment Using Public Datasets with Talend + AWS

Consisting of a series of randomized, placebo-controlled trials on tumor and cancer patients, the data was collected in a span of 3-6 months at various clinics and hospitals across the United States. The data was then organized by several dimension tables such as the names and addresses of the treatment centers, along with factual test- related information.

Data Ingestion

Talend Big Data Platform downloads the files from the external site hosting the clinical trial data to an FTP server running on an AWS EC2 cluster. The files on the FTP server are then sent through the Internet Gateway to an AWS VPC (Virtual Private Cloud) environment.

Flattening XML Files

Talend Big Data Platform retrieves hierarchical XML data and flattens it to generate one fact table and multiple dimensional tables leveraging the Talend Data Mapper feature in the platform to perform the mappings. Talend Data Mapper is used to transform complex hierarchical data (for example, nested or looping structures). It lets you map between data records or documents in various formats, such as XML, SWIFT, COBOL, CSV, EDI, XLS, and more.

Each clinical XML file contained information about the study being conducted as well as various dimensional attributes such as clinical locations and types of drugs being tested. The challenge here was to parse out necessary dimensional and factual data from hierarchical XML files, which at times meant one-to-many records per XML file. The Talend Data Mapper was utilized to effectively flatten out complex XML structures into multiple fact and dimensional tables.

Data Staging and Target Data Warehouse

After transforming the XML data, Talend Big Data Platform then bulk loads the files to an Amazon S3 bucket for staging and then once again, bulk loads it in parallel to Amazon Redshift, a fast and powerful, fully managed, petabyte-scale data warehouse, for reporting and analysis. In the end, Talend transformed a batch of XML files into as many as 52,000 records into the fact table and up to 200,000 records into dimensional tables.

PRODUCTS IN THIS ARCHITECTURE

- Talend Big Data Platform
- Amazon Virtual Private Cloud (VPC)
- Amazon S3
- Amazon Redshift
- Amazon Auto-scaling Groups

USE CASE 3: ENABLING SOCIAL AND MOBILE ANALYTICS TO IMPROVE MARKETING FOR A FOOD & BEVERAGE SERVICE RETAILER

With a need to design better targeted marketing campaigns and to increase its brand awareness, a leading food and beverage service retailer wanted to look beyond its store data and into its mobile app and social media channels for better customer insights. However, their legacy ETL system limited their business users' ability to do so.

Prioritizing a 100% cloud infrastructure for speed, flexibility, and scalability, Talend Real- Time Big Data Platform, together with AWS S3, EMR, and Redshift was used for a proof-of-concept to perform predictive analysis in a

secure AWS cloud environment. With that, they were able to analyze social data from a sample set of 10 million app users.

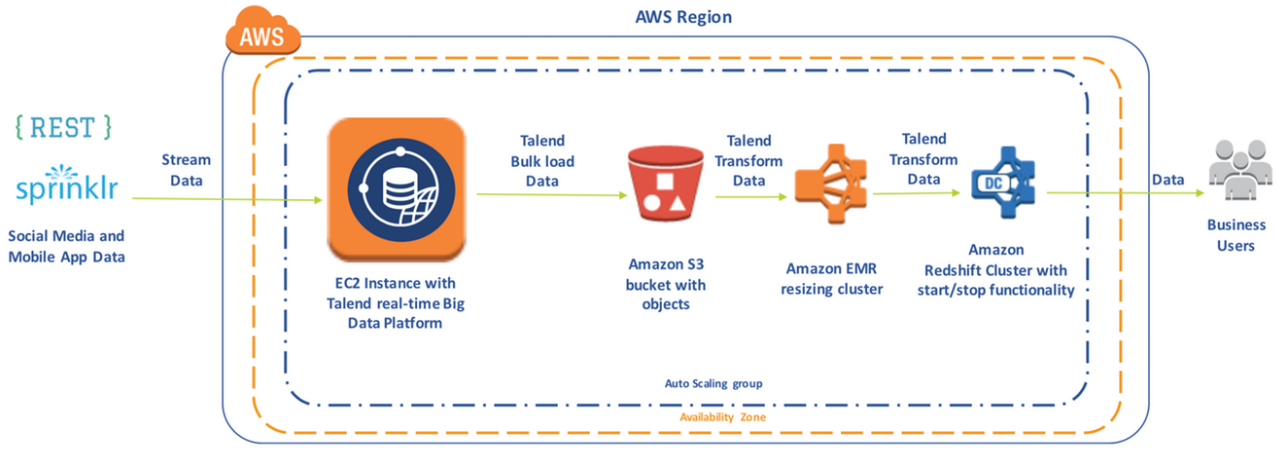

Figure 3: Social and Mobile App Data Analytics Infrastructure with Talend + AWS

The retailer used Sprinklr, a social media management platform that collects and analyzes data from a broad range of social media channels to discover their customers' sentiment about their business, and ultimately use the analysis to shape and redesign their social media content to better engage with them.

Data Ingestion

The Talend Real-Time Big Data Platform ingests the raw data in JSON format through the Sprinklr REST API using the appropriate authentication parameters. It then converts the data into flat files and filters only the information needed.

Data Transformation and Analytics

Talend Real-Time Big Data Platform then bulk loads the transformed files into AWS S3. In AWS S3, the Talend job server creates control parameters to filter data based on directories in a specified bucket. A framework is built within the Talend platform to handle a specific directory structure allowing for the replay of data into AWS EMR and subsequently into Redshift.

Talend supports automatic cluster resizing for EMR and Redshift so that customers can optimize their computing and storage resources by dynamically changing the number of nodes according to their workloads. Talend also enables Redshift clusters to start and stop automatically when jobs are ready to run or complete, which is useful for bulk load processes that only run periodically. It also eliminates the need to write estimate scripts that start the Redshift cluster.

Security on AWS

Security was a top concern for the retailer, therefore they leveraged Talend's support for both AWS S3 server and client-side encryption. S3 server-side encryption protects data at rest, while client-side encryption makes sure the data is protected before it gets sent to its S3 destination.

PRODUCTS IN THIS ARCHITECTURE

- Talend Real-Time Big Data Integration
- Talend Data Preparation
- Amazon S3
- AWS Redshift
- AWS EMR
