Building a Big Data Architecture on AWS To Understand Customer Preferences

Business problem: Understanding Customer Preferences

Businesses in multiple verticals have a need to understand their customers' preferences and interactions with their products and services so that they can proactively make relevant offers to their users.

Some examples of businesses and systems where this type of need exists include:

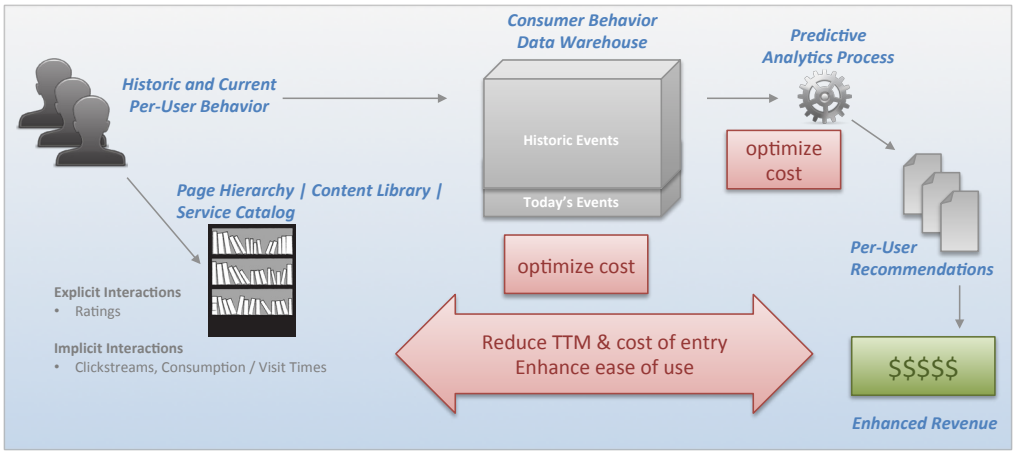
- Media & Entertainment: Connected TV platforms / "Over-the-Top" content distribution platforms
- Retail: Online retailers, consumer-facing websites, clickstream analysis and abandoned carts
- Hospitality: Loyalty programs spanning diverse properties and service offerings

To create accurate and timely per-user recommendations, companies typically need to aggregate data regarding interactions between users and each of the company's products and services. This data includes each user's historic behaviors, offer interactions and responses, consumption, stated or observed preferences, and provided ratings and feedback. This information can be captured in a data warehouse, which then makes the data available for downstream analytics processes, which prepare recommendations and offers for each user.

The desired per-user recommendations are typically produced by a predictive analytics, machine learning, or collaborative filtering algorithm. When the number of users, products, and interactions is large, processing these recommendations requires the use of a distributed computing cluster such as Hadoop. The Hadoop ecosystem provides an increasingly mature portfolio of offerings capable of filling this role.

Data warehousing solutions combined with a big data analytics infrastructure built for the Hadoop ecosystem provide powerful capabilities for optimizing interactions, offers, and recommendations for customers in real-time. However, for many companies these solutions are out of the reach of line-of-business owners: the cost and lead-time requirements are too high, due to the need for a large up-front infrastructure investment, along with the need to hire expensive big data experts.

The diagram below summarizes these cost challenges encountered in analyzing this data.



800 Bridge Parkway Suite 200, Redwood City, CA 94065
www.talend.com | info@talend.com

6525 Gunpark Dr Ste 370-168, Boulder, CO 80301
www.47lining.com | info@47lining.com

Commissioning a new data warehouse is expensive, requires large teams with specialized skillsets, and long lead times to get the data warehouse up and running. In cases where a data warehouse already exists within the enterprise, obtaining approvals and resources required to use it to address problems like the one described above can be very difficult.

Commissioning a big data analytics initiative can similarly be very expensive. Using Hadoop has traditionally required strong support from third-party vendors, coupled with a team of data scientists who are familiar with the wide range of tools in the Hadoop ecosystem and what is required to set them up and use them effectively.

How can we reduce the time to demonstrate integrated capabilities, enhance ease of use, and optimize costs for big data systems?

Decreasing the cost of big data deployments

A cost-effective solution combines the rich library of data transformations provided by Talend Studio and Integration Cloud with **on-demand**, elastically priced data warehouse and big data services from Amazon Web Services (AWS).

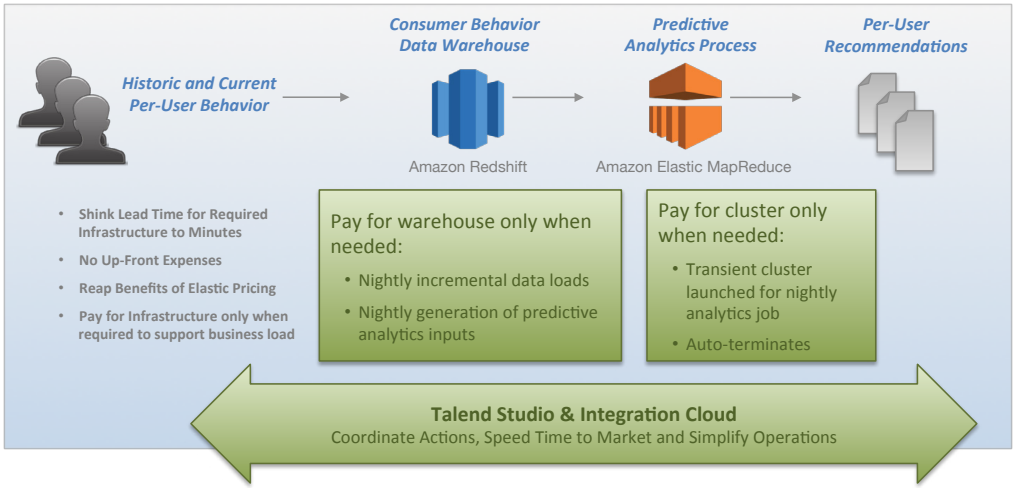
AWS' "on-demand" infrastructure can be deployed in minutes whenever it is needed to perform a workload. When the workload is complete, the infrastructure can be released, which "stops the clock" on payment for that infrastructure.

This is called **elastic pricing**: you pay only for what you use. This means that pricing tracks closely with your actual consumption. This is in contrast to traditional systems, which must be sized to accommodate the peak load that must be supported. Elastic pricing for an on-demand infrastructure can result in substantial cost savings.

AWS provides a set of related capabilities that make this solution approach possible:

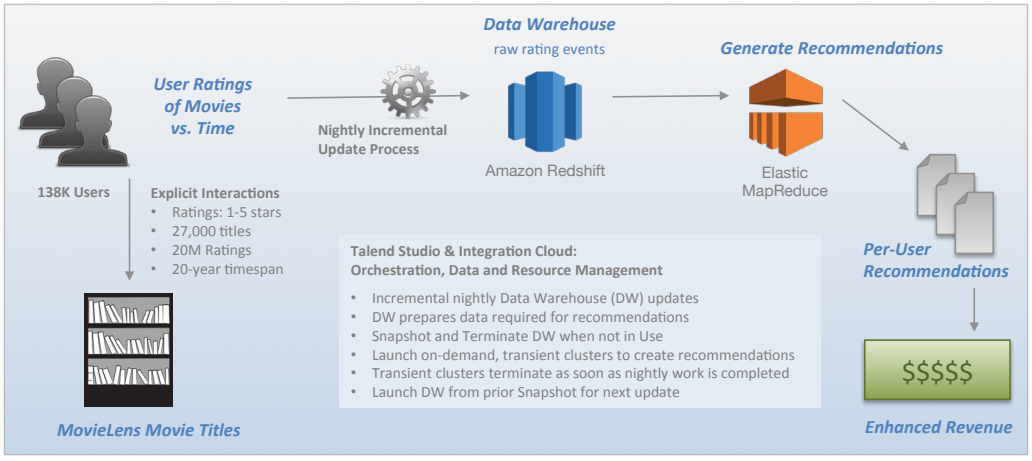
- ***Amazon Redshift*** is a cloud-based data warehouse environment that can be provisioned on demand.
- ***Amazon Simple Storage Service (S3)*** provides an inexpensive, highly durable, scalable, distributed object store. Data in S3 can persist independently from Redshift or Elastic MapReduce resources (see below). Because S3, Redshift, and EMR are all cluster-oriented distributed computing systems, data loads from S3, snapshots to S3, and restores from S3 are scalable as data and cluster sizes grow.
- ***Redshift Copy from S3*** – Redshift provides for extremely performant parallel loads of data from S3 by distributing data loading work across all nodes in the Redshift cluster and leveraging high-bandwidth I/O between each Redshift cluster node and compute nodes serving S3 data.
- ***Redshift Snapshot / Restore from Snapshot*** – Redshift can efficiently capture its current state in S3. Snapshots created prior to cluster termination can be used to easily launch replacement cluster(s) as needed in the future.
- ***Transient Elastic MapReduce (EMR) Clusters*** – In a traditional Hadoop cluster, the cluster serves two roles: to process work, and to persist input and output data by contributing local storage for use within the cluster's Hadoop distributed file system (HDFS). Amazon EMR includes drivers that enable each node in the EMR cluster to efficiently read and write directly from S3, removing the need to stage data in HDFS. When HDFS is not required, cluster nodes only need to exist during the time that work is being processed. This is called a ***Transient Cluster***: one that is launched when needed, pulls its required input data from S3, persists its desired output data in S3, and then terminates. You pay for the cluster only while it is running.

Talend provides an unparalleled set of capabilities to marshal, transform, and integrate many data sources. For this reason, Talend fills a critical need to reduce compute and resource costs in many organizations: it de-risks the data management project and allows integration specialists who don't have an extensive background in Hadoop to operate and configure big data workflows. As shown below, Talend makes it easy to schedule and orchestrate work, and is a perfect fit for deploying and managing the on-demand and elastically priced resources provided by AWS. In addition, Talend Integration Cloud makes the underlying integration workflows and results easily accessible to other collaborators in diverse roles throughout the organization.



Solution Architecture Overview

The solution architecture shown below demonstrates the benefits of using Talend to access elastically priced, on-demand big data resources provided by AWS.



In this scenario, we first establish an on-demand data warehouse based on Amazon Redshift to contain raw rating events and related movie title information. We will use data from GroupLens, a research group in the Department of Computer Science and Engineering at the University of Minnesota. GroupLens provides the MovieLens ml-20m dataset, which includes 20,000,263 user ratings across 27,278 movies. This data was created by 138,493 users between January 09, 1995 and March 31, 2015.

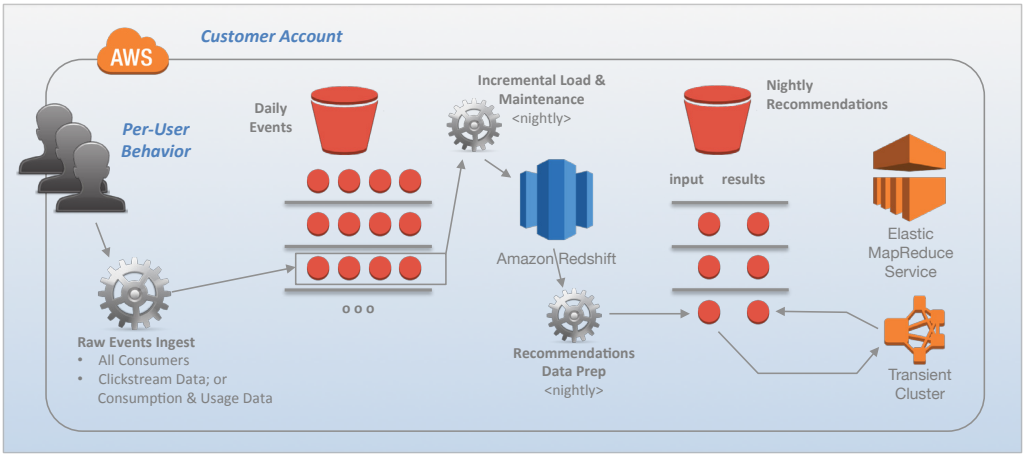
Rating event data for the prior day is loaded incrementally each night to the Redshift-based data warehouse. Then the data warehouse will create a transformation that includes a specified subset of all historical data, including the most recent incremental load, suitable for use by a recommendations generator. When the work of loading incremental events and preparing input for the recommender is done, we'll snapshot the Redshift cluster and terminate it to achieve the benefits of elastic pricing. We'll reconstitute the cluster from the generated snapshot the following evening.

The recommender will be based on a transient Elastic MapReduce cluster which will only be deployed when recommendations work needs to be performed.

We will show how Talend Studio is used to represent each part of the overall flow, and how Talend Integration Cloud orchestrates data movement and management of AWS on-demand resources as the process runs each night.

Implementing the solution architecture with Talend and AWS

The diagram below depicts the AWS data and services flow within the demonstration.



Specific steps within this flow are as follows:

Raw events ingest. A public S3 Bucket includes a demonstration subset of the MovieLens dataset that has been organized by the date of each movie rating by a user. Data within the bucket is organized in a hierarchy based on year, month, and day. In the demonstration example, incremental raw events for each day are placed in the appropriate S3 location by code that is monitoring application logs and generating rating events. Importantly, the data warehouse is not required to be available for this data collection step to occur.

Daily events in S3. Each day's events are placed in S3 in the format required for them to be seamlessly ingested into Amazon Redshift via its parallel, performant "COPY FROM S3" capability.

Incremental data load and cluster maintenance. Each night, a single COPY FROM S3 with the correct key glob pattern will pull in and incorporate all of the prior day's rating events into the data warehouse. Data load time is minimized because data will be loaded in parallel across all nodes in the cluster.

Recommendations input data preparation. The Redshift-based data warehouse is very efficient at creating summaries and data transformations required to prepare input data as required by the recommendations-generation process. These data are generated in a temporary table in redshift then offloaded in parallel to S3 using the UNLOAD TO S3 capability. An S3 Bucket is used to persist each day's input to and output from the recommendations process. Data within the bucket is organized in a hierarchy based on year, month, and day.

Redshift cluster snapshot and termination. Once incremental data has been ingested and the prior day's historical summary has been prepared and unloaded to S3 for use by the recommendations process, a snapshot of the Redshift cluster is created, and the cluster is terminated until the following night's activity.

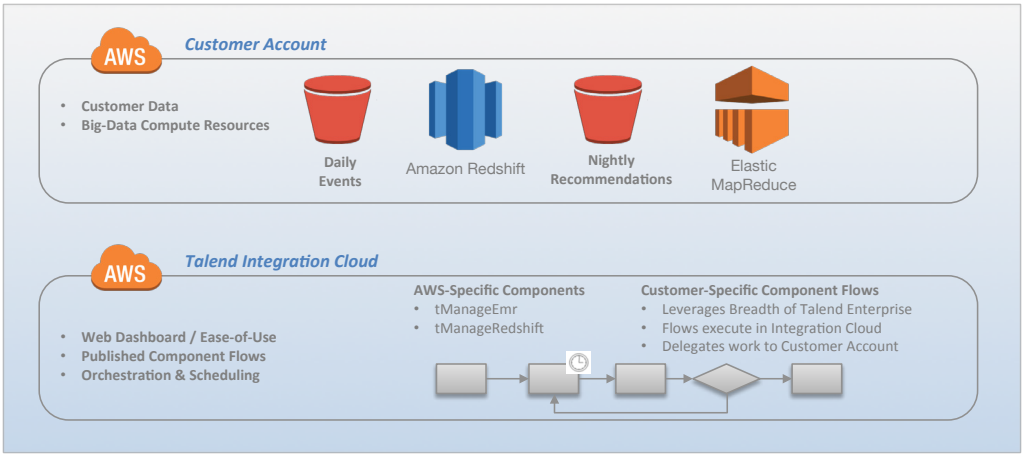
Recommendations generation. A transient EMR cluster is commissioned to generate recommendations from the historical ratings information that has been prepared in S3 for this purpose. The cluster draws its input directly from S3, so no orchestration of data movement into hdfs is required.

Recommendations output. The cluster writes its recommendations output directly to S3 so that recommendations remain available to downstream consuming processes even though the EMR cluster that generated them is no longer in service. Once recommendations are generated, the transient cluster is terminated.

Orchestration with Talend Studio and Integration Cloud

Use of Talend Studio and Integration Cloud enables ease-of-use and partitioning of resources between Talend's Integration Cloud – a hosted platform for scheduling and managing end-to-end big data flows – and the Customer's AWS account. Customer data and commissioned big data components such as the Redshift-based data warehouse and Elastic MapReduce clusters reside in the customer's AWS account.

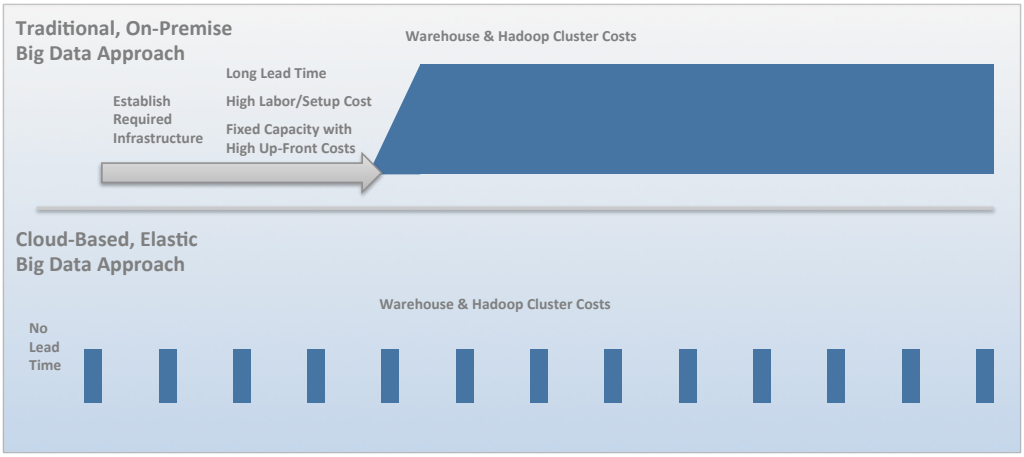
Talend Studio is used to create low-level component flows consistent with the requirements of the demo, using AWS-specific components that enable management of on-demand AWS resources to achieve elastic pricing. These customer-specific flows are published to Talend Integration Cloud, where high-level end-to-end flows are scheduled and orchestrated.



Benefits

Using this approach, the benefits of on-demand big data resources and elastic pricing are considerable:

- Drastically shrinks lead time and required budgets to establish business-relevant big data capabilities
- As shown in the figure below, this approach uses transient resources commissioned on-demand; you pay only for what you use to support business needs
- Talend provides ease-of-use and accessibility of big-data flows to contributors throughout your organization
- Enhanced AWS-specific components in Talend 6.1 support on-demand management and elastic pricing



800 Bridge Parkway Suite 200, Redwood City, CA 94065
www.talend.com | info@talend.com

6525 Gunpark Dr Ste 370-168, Boulder, CO 80301
www.47lining.com | info@47lining.com