Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

Data Processing Pipelines in Business

The Amazon Web Services (AWS) Data Pipeline is a popular manifestation of the data pipeline found in business today. AWS Data Pipeline can be seen as a tool in a software programmer's toolbox. It is a web service that a programmer uses to automate the movement and transformation of data. With AWS Data Pipeline, a programmer can define data-driven workflows, so that tasks can be dependent on the successful completion of previous tasks. The programmer defines the parameters (inputs/outputs) of their data transformations and the AWS Data Pipeline enforces the logic that they have set up. A description of the AWS Data Pipeline can be found here:
http://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html

1. Cloud computing gives a company an opportunity to run services for customers "in the cloud" instead of within the company's own IT department. One advantage of using the cloud is that the cloud can expand to an increasing number of customers coming to a service much better than a company's IT department can. Suppose a company chose to create a new service for its customers and was deciding whether to host it in the cloud or within their IT department. How might the AWS Data Pipeline fit into their choice to use cloud computing?

The choice of using the AWS Data Pipeline would go hand-in-hand with the choice to use cloud computing. Per the information about the AWS product from the above link, by choosing the AWS Data Pipeline solution the company would acquire access to the Amazon cloud and it's scalable processing and storage capabilities. Specifically, the company would have access to four Amazon services to store data (DynamoDB, which is a noSql database; RDS, which is a relational database; Redshift, which is a data warehouse; and S3, which is a tool for object storage, such as, log files) and two services to process/transform data (EC2, which stands for Elastic Compute Cloud; and EMR, which stands for Elastic Map Reduce). All of these AWS Data Pipeline services are scalable, so as the company acquires new clients, the Amazon cloud should be able to accommodate additional workload and storage requirements.

The company would also have control over the data pipeline process flow, as that is defined as part of the AWS Data Pipeline "pipeline specification" along with task schedule definitions. Amazon offers several tools the company would be able to use to access and manage the pipeline.

The AWS Data Pipeline option is also consistent with the cloud pricing model, which is "you pay for what you use".

The only other issue to consider is data transmission over the public internet. If the service the company is offering includes collecting large volumes of data in local servers, the company will need to evaluate potential data throughput issues as it transmits data to the Amazon cloud. That will be particularly critical if the service the company is offering to its customers require processing collected data immediately after data collection.

2. Use the Internet to research actual uses of the Data Pipeline. Find a couple of uses. For each, discuss in what kind of business is the Data Pipeline employed and how the pipeline is employed (for what use).

There are many success stories at the Amazon AWS site "Case Studies & Customer Success Stories" [1] describing successful uses of the AWS Data Pipeline solution. The one from NASA [2] caught my attention. NASA wanted to allow users all over the world to see video (live video streaming) of its Curiosity rover landing on Mars. But it couldn't know how many people would want to watch the images online at any given time, or where these people would be, so it used Amazon servers all over the world, on an on-demand basis, to make the video available across the globe as users tried to watch it online.

An interesting success store for Microsoft Azure, a competitor to Amazon AWS, comes from a solution developed for NBC [4]. NBC wanted to "push" news notifications to its mobile clients very quickly, ahead of the competition. They used a MS Azure service called "Windows Azure Notification Hubs" [5] which is a solution designed exactly for what NBC needed.

I found two examples of data pipelines that include analytics. The first comes from Google Cloud Compute. It is a solution for mobile gaming [6] that employs the Dataflow programming model. The pipeline collects data from users playing online and performs analytics to display back to the gamers useful information about their performance.

A second example is a solution developed by IBM for the Olympic US cycling team, specifically for the "team pursuit" competition [7]. The solution [8] uses sensors, including wearable sensors, power meters and health data monitors to collect data, and employs real time data transmission to the cloud, where analytics is performed on the data, and results are transmitted back to the coaches' tablets, for decision making at individual and team level, on the fly.

[1] https://aws.amazon.com/solutions/case-studies/

[2] https://aws.amazon.com/solutions/case-studies/nasa-jpl-curiosity/?pg=main-customer-success-page

[3] https://www.quora.com/Who-are-AWS-major-competitors-by-market

[4] https://customers.microsoft.com/Pages/CustomerStory.aspx?recid=13325

[5] https://azure.microsoft.com/en-us/documentation/services/notification-hubs/

[6] https://cloud.google.com/dataflow/examples/gaming-example

[7] http://www.sporttechie.com/2016/03/28/analytics/ibm-helping-u-s-womens-cycling-team-prepare-ride-past-competition-2016-rio-olympics/

[8] http://www-01.ibm.com/software/ebusiness/jstart/index.html