Database Host on AWS

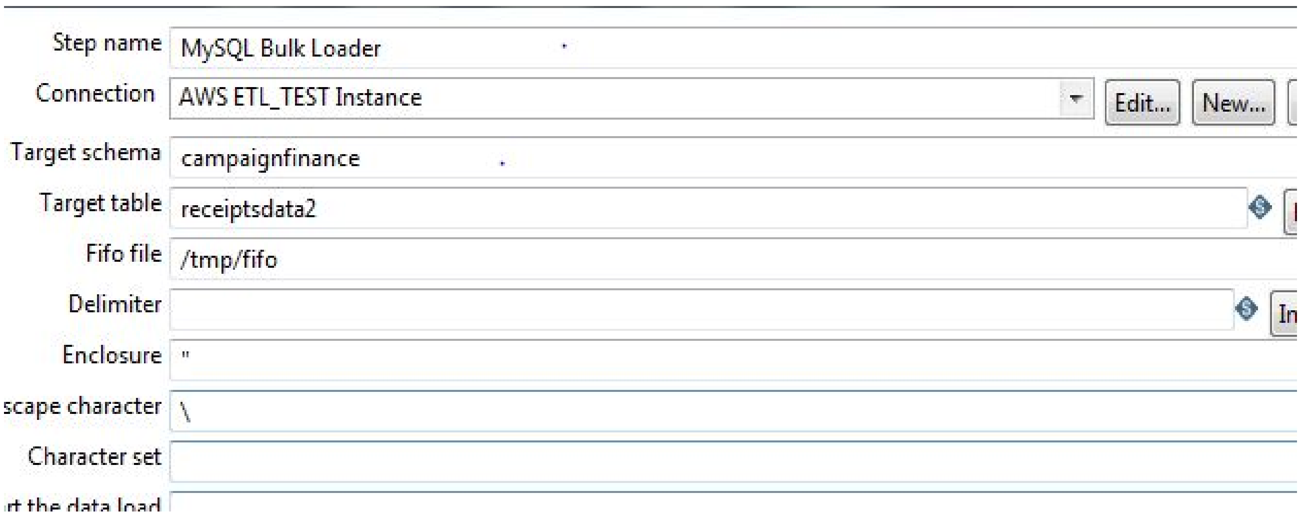
pentaho

ETL

amazon web services™

Database Questions

1. MySQL Bulk Loader step in Pentaho: We are having issues with the Fifo file parameter. Can use this step when Spoon is installed on a Windows machine? We are running Pentaho locally and piping the data into AWS. Please see below screen capture.



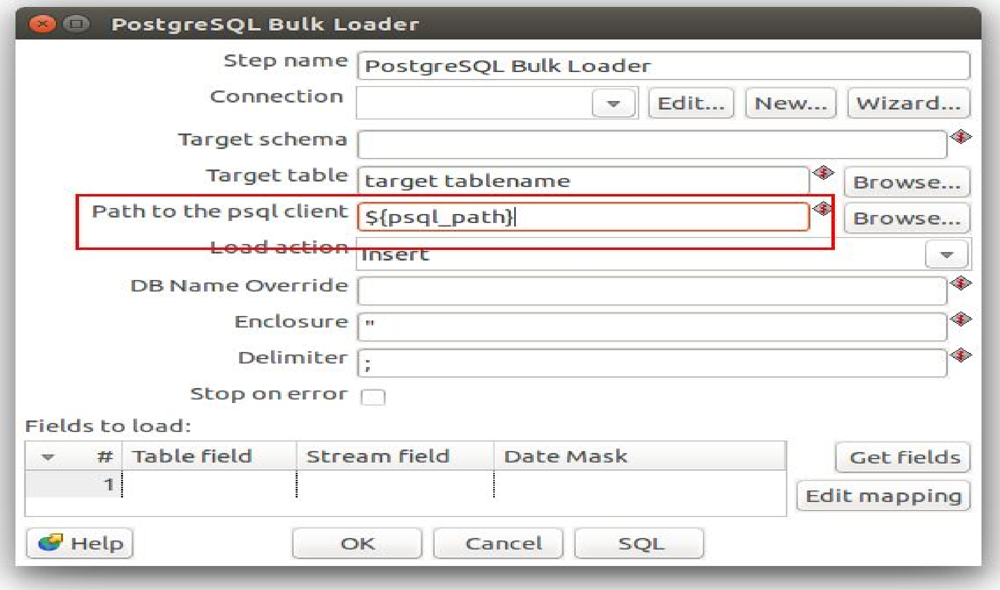
Step name: MySQL Bulk Loader
Connection: AWS ETL_TEST Instance
Target schema: campaignfinance
Target table: receiptsdata2
Fifo file: /tmp/fifo
Delimiter:
Enclosure: "
scape character: \
Character set:
rt the data load

ANS:- Fifo File - This is the fifo file used as a named pipe. When it does not exist it will be created with the command mkfifo and chmod 666 (this is the reason why it is not working in Windows).
Circumvention: Use the MySQL bulk loader job entry to process a whole file (suboptimal)
Not supported but worth to test: mkfifo and chmod are supported by the GNU Core Utilities

2. We received an error message relating to the 'path to the psql client' in PostgreSQL bulk loader step. How can we find and apply the path to the psql client on our amazon EC2 instance running PostgreSQL?

ANS:- First we need to define a parameter called psql_path in kettle.properties file.
E.g. psql_path=c\:/Program Files (x86)/pgAdmin III/1.16/psql.exe

Then we need to set Bulk Loader's "Path to the psql client" property, we can use ${psql_path}



PostgreSQL Bulk Loader
Step name: PostgreSQL Bulk Loader
Connection:
Target schema:
Target table: target tablename
Path to the psql client: ${psql_path}
Load action: insert
DB Name Override:
Enclosure: "
Delimiter: ;
Stop on error: □

Fields to load:

#	Table field	Stream field	Date Mask
1			

3- What parameters should be set to increase data transfer speeds to a postgres database?

ANS:- Performance PostgreSQL
optimization settings of PostgreSQL, depend not only on the hardware configuration, but also on the size of the database, the number of clients and the complexity of queries, so that optimally configure the database can only be given all these parameters.
PostgreSQL settings (add/modify this settings in postgresql.conf and restart database):

1. max_connections = 10
2. shared_buffers = 2560MB
3. effective_cache_size = 7680MB
4. work_mem = 256MB
5. maintenance_work_mem = 640MB
6. min_wal_size = 1GB
7. max_wal_size = 2GB
8. checkpoint_completion_target = 0.7
9. wal_buffers = 16MB
10. default_statistics_target = 100

4:- If there are unallowed characters for postgres text field, what is the best way to handle those; ie ASCI Null?

Ans:- solution 1:- The NULLIF function returns a null value if value1 equals value2; otherwise it returns value1. This can be used to perform the inverse operation of the COALESCE

solution2:- There are different ways to handle special characters.
E.g. Escaping single quotes ' by doubling them up ->''is the standard way and works of course.
E.g. ~~`'user's log'`~~ `'user''s log'`

5:- We are moving data from a DB2 database to AWS. The goal is to update the data in less than 8 hours. We have nine tables and the largest table includes about 130 million rows. Is this feasible? What is the best way to implement this strategy on AWS?

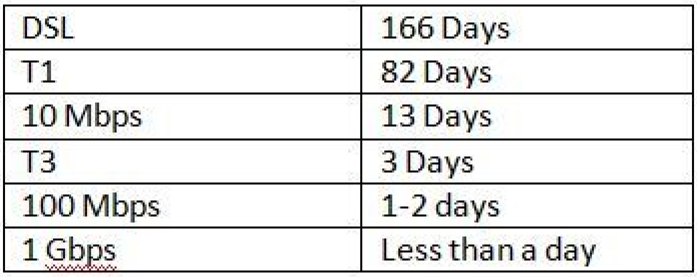
Ans:-solution 1:- In the first two parts of this series we discussed two popular products--out of many possible solutions--for moving big data into the cloud: Tsunami UDP and Data Expedition's ExpeDat S3 Gateway. Today we'll look at another option that takes a different approach: Signiant Flight.

solution 2:- AWS Import/Export is a service you can use to transfer large amounts of data from physical storage devices into AWS. You mail your portable storage devices to AWS and AWS Import/Export transfers data directly off of your storage devices using Amazon's high-speed internal network. Your data load typically begins the next business day after your storage device arrives at AWS. After the data export or import completes, we return your storage device. For large data sets, AWS data transfer can be significantly faster than Internet transfer and more cost effective than upgrading your connectivity

solution 3:- Snowball is a petabyte-scale data transport solution that uses secure appliances to transfer large amounts of data into and out of the AWS cloud. Using Snowball addresses common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns. Transferring data with Snowball is simple, fast, secure, and can be as little as one-fifth the cost of high-speed Internet.

6:-What is the largest dataset (relational database table) that Pragmatic has moved to AWS? How long did it to update such a table? What performance strategies did Pragmatic undertake to achieve peak performance for updating such a table?

Ans:- If you look at typical network speeds and how long it would take to move a terabyte dataset:



DSL	166 Days
T1	82 Days
10 Mbps	13 Days
T3	3 Days
100 Mbps	1-2 days
1 Gbps	Less than a day

Depending on the network throughput available to you and the data set size it may take rather long to move your data into Amazon S3. To help customers move their large data sets into Amazon S3 faster, we offer them the ability to do this over Amazon's internal high-speed network using AWS Import/Export.

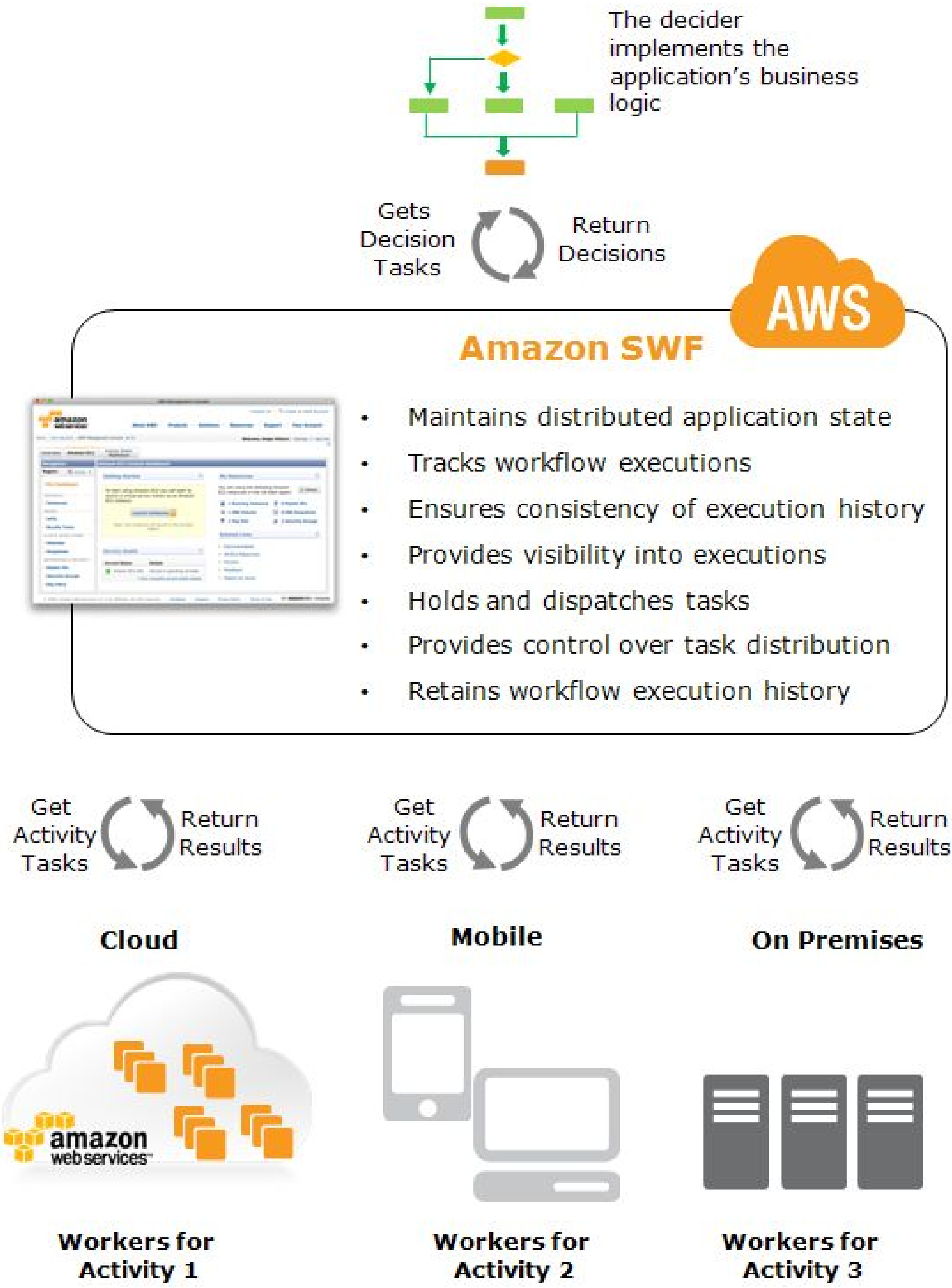
7:- What is Pragmatic suggested approach for setting up ETL architecture for an AWS based datacenter?

Ans:- With Amazon Simple Workflow (Amazon SWF), AWS Data Pipeline, and, AWS Lambda, you can build analytic solutions that are automated, repeatable, scalable, and reliable. In this post, I show you how to use these services to migrate and scale an on-premises data analytics workload.

Workflow basics

A business process can be represented as a workflow. Applications often incorporate a workflow as steps that must take place in a predefined order, with opportunities to adjust the flow of information based on certain decisions or special cases.
The following is an example of an ETL workflow:

The graphic below is an overview of how SWF operates.



- Maintains distributed application state
- Tracks workflow executions
- Ensures consistency of execution history
- Provides visibility into executions
- Holds and dispatches tasks
- Provides control over task distribution
- Retains workflow execution history

8:-Rather than using Pentaho CE for ETL and reporting, what do you think are the advantages/disadvantages of implementing a hybrid environment running Pentaho ETL and Tableau Server? Have you implemented such a mixed environment for any of your clients?

Ans:- This can be done. Tableau does not have ETL. So we can use Pentaho ETL with Tableau. We have worked in combination with Tableau and Pentaho.You can use Pentaho for ETL & visualize data using tableau.

9:- Do you have any clients in the United States that use Pragmatic support for Pentaho?

Ans:- We are a products and services company working primarily in ERP< CRM, BI and Analytics. We have worked with several customers from United States and can give you a reference for ERP deployment and report generation.

10:-Do you have any clients in the United States that used Pragmatic consulting services for setting up their ETL architecture? If so, do you mind listing them as a referral?

Ans:- We have customers who have used our AWS consulting expertise not only limited to Pentaho, in the United States but in entire world in countries such as Australia,New Zealand, Switzerland, Belgium. We have also deployed scalable architectures on AWS cloud.But unfortunately most of these are companies are middle men and since we have signed NDA with them, we cannot declare their names. But we can definitely give you reference of companies in United STates with whom we have worked with other technologies such as ERP. Will that work for you?