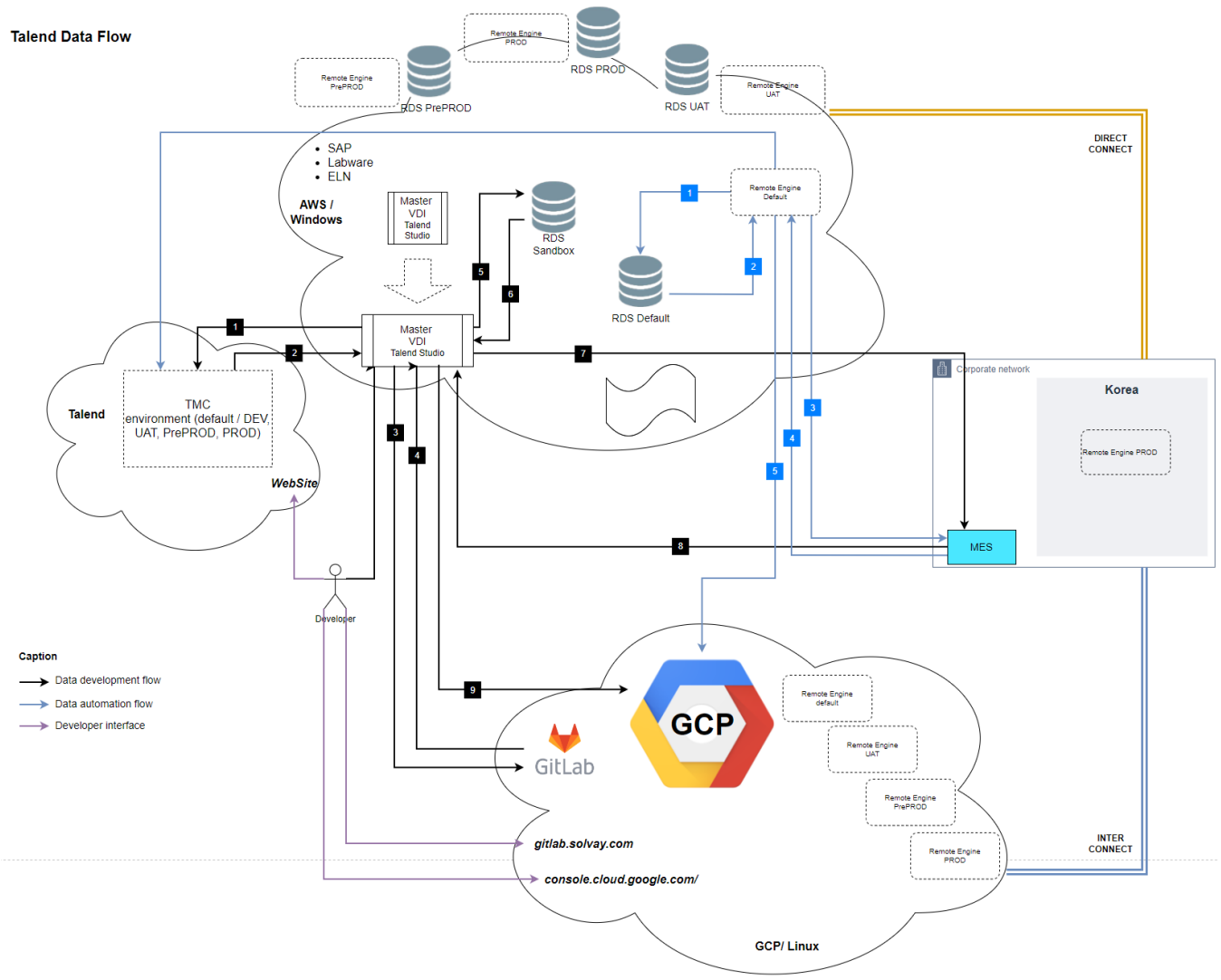
Talend Architecture



Schema original file :

Talend architecture is an hybrid platform between AWS, GCP and solvay onPremise machines. The platform is not designed to take into account confidential or sensitive project such as EAR, ITAR, GDPR topics. In case confidential or sentitive projects need to be done on talend, a new project should be set to assess the use cases, design a specific talend architecture which fit to solvay security policy and standard.

Remote Engines

There are for the moment (August 2022) 9 remote engines running as following :

- 4 remote engines on AWS (DEV / UAT / PRE-PROD / PROD) => all machines on AWS are using windows server 2016 datacenter edition
- 4 remote engines on GCP (DEV / UAT / PRE-PROD / PROD) => all machines on GCP are using ubuntu 20.04 LTS edition
- 1 remote engines on EHWA (Korea) Solvay plant (PROD) => the machine is running under windows server 2016 standard edition, normally to be aligned with the rest of the architecture we should get as well 3 more machines to build DEV / UAT and PRE-PROD but due to talend license contrainst there is a maximum of 10 remote engines available. The last remote engine installation available is kept for other incoming use case not defined yet ...

Initially the old platform was only running on windows machines but according to internal solvay policy, it is asked to run as much as possible on Linux machines. However due to technical / policy constrainst it is not possible to get windows machine within GCP Cloud provider that's why all windows machines are located within AWS. Windows machine is required for Industrial talend project which use specific and proprietary driver to connect on MES (OSISOFT PI) system which is only availble for windows machine. Linux machines are set on GCP because 95% of the target systems use cases for data loading is located in GCP cloud provider.

The remote engine located on Korea server is for a very specific use case to retrieve Korea (EHWA) RnI Battery cycler data.

To sum-up having remote engines located on AWS, GCP and onPremise should help to anticipate further use cases depending the source/target systems location.

Remote engine usage :

By default GCP Linux engine must be used except for project which required windows specific components such as MES Proprietary drivers or windows shared drive.
Remote engine selection should also be made based on performance matters, it is recommended to perform some connectivity and test performance before choosing the right remote engine for a project.
Once the type (GCP or AWS) of remote engine is selected, it is suitable to keep it for the whole project in order to avoid duplicate folders / files and so on.... since there is no global shared folder available among the remote engines for the moment.
Having a global shared folder could be made later on depending the needs such as remote engine clustering for example. For the moment talend project folder must be stored on remote engine machine.

Connection on AWS could only be done with ADMS account + yubikey. Connection on GCP are managed through google groups gcp-talend-developers-nonprod@solvay.com and gcp-talend-developers-prod@solvay.com managed by Data OPS team.

In the long term windows machines should be decomissionned.

RDS Databases

The RDS databases are required in order to avoid having talend context variable (for example connection string to data source system, credendtials, reference date and so on) hardcoded within talend project and to ease the run of talend project theses talend parameters were stored on external database such as AWS RDS. Parameters could be changed without having talend studio application or re-deploy talend jobs again.

There are 5 RDS Databases available on AWS cloud provider : DEV / UAT / PRE-PROD and PROD, the 5th environments is the SandBox dedicated to talend developer for running locally their jobs from their computer / VDI (Virtual Desktop Infrastructure) / machine.

The RDS databases beside storing context variable values are also storing talend project technical (the technical logs generated during a job run) and functional (the logs set by the developers / data engineers) logs.

Gitlab

Talend development files are versionned on solvay gitlab repository. The access to gitlab through Talend is made with a SSH Key specific to each developer / data engineer.

TMC

The administration of the platform is made through the TMC hosted by talend editor (AWS).

Network

The connection between AWS / GCP and Solvay onPrem is made with Direct connect (Solvay to AWS) and Inter connect (Solvay to GCP), globally GCP / AWS and Solvay network is considered as one. the TMC is considered as a third party despite it is hosted eventually on AWS by Talend editor.

All talend internal connection are made with SSL encryption :

- Talend Studio to TMC
- TMC to Remote engine
- Talend studio to remote engine (DEV only)
- Talend studio to RDS
- Remote engine to RDS
- Except Talend studio to gitlab is going through SSH

WARNING :

Direct connect is a private network managed by BT / Equinix and it is not encrypted by default, All connections made with talend MUST BE ENCRYPTED in order to ensure **an end to end encryption** as per solvay security policy.

Inter connect is only available for VM and Cloud SQL GCP Services, by default others GCP SaAs services are open to public (meaning GCP services can be reach from solvay outside network) such as GCS or GBQ.

Global Data Flow

On Talend platform there are two distincts data flow, the one during the project development and the schedule one (automation).

Development stage

1. When the developer / data engineer is working on talend studio a first connection is made to the TMC
2. The TMC returns the developer / data engineer project access rights
3. Once the project is selected in talend studio, a connection is made on gitlab
4. Once connected on Gitlab, the files are retrieved in a local branch within talend studio for the project identified

5. As soon as a job is run within talend studio a connection is made to the RDS database
6. Once connected to the RDS database, talend context parameters values are sent back
7. With the correct connection string and credentials, a connection is made on the source system (on the picture : MES)
8. The data are retrieved in talend studio
9. When the data are processed there are mainly sent to GCP (mainly GCS and GBQ)

Scheduled stage

Once talend jobs are deployed and running

1. The job deployed on a remote engine will connect to the RDS database
2. Retrieve the context parameters values
3. Connect on the source system
4. Retrieve the data from the system
5. Once the data are processed, It is stored on GCP (mainly GCS and GBQ)

Architecture enhancement suggested

by importance order

1. During the platform installation, due to budget / priority constraints and security policies, by default remote engine logs are not supposed to be send to the TMC. A Monitoring is needed
2. Due to prioritization, Talend studio are installed on each developer / data engineer machine locally, using VDI or Virtual environment would be very helpful to onboard newcomers and centralized development application update/upgrade and fix the version used among the developers / data engineers
3. Add a common shared folder available for all remote engines, it will permit to centralize the data and it is a prerequisite if fail over or clustering need to be set up.
4. CI/CD implementation, actually promotion from one environment to another one is made manually through the TMC. CI / CD pipeline could be set to automatize the promotion.
5. Centralize talend external jar/components and studio upgrade by setting a Nexus server. Going on talend cloud there is no Nexus to manage but it prevent to centralize third party components / libraries and talend studio update / upgrade
6. Dockerised remote engines, actually remote engine are set on virtual machine, in order to manage the horizontal scale it could be good to use containezization solution
7. Set a direct connection between AWS and GCP to improve network route
8. According to DEV Remote engine usage check if it is possible to use Solvay Root Certificate