WHAT IS
APACHE AIRFLOW ?
(GOOGLE CLOUD COMPOSER)

By Guha Rajan

What is Apache airflow / cloud composer

- Apache Airflow is a workflow automation and scheduling system that can be used to author and manage data pipelines. Airflow uses workflows made of directed acyclic graphs (DAGs) of tasks.

- As it's a open source product, google cloud has included it in their cloud platform, hence it does not require any install of this product locally.

  - You can get free access to Google (https://cloud.google.com) requires credit card authentication. **Please refer Google site for Terms and conditions**

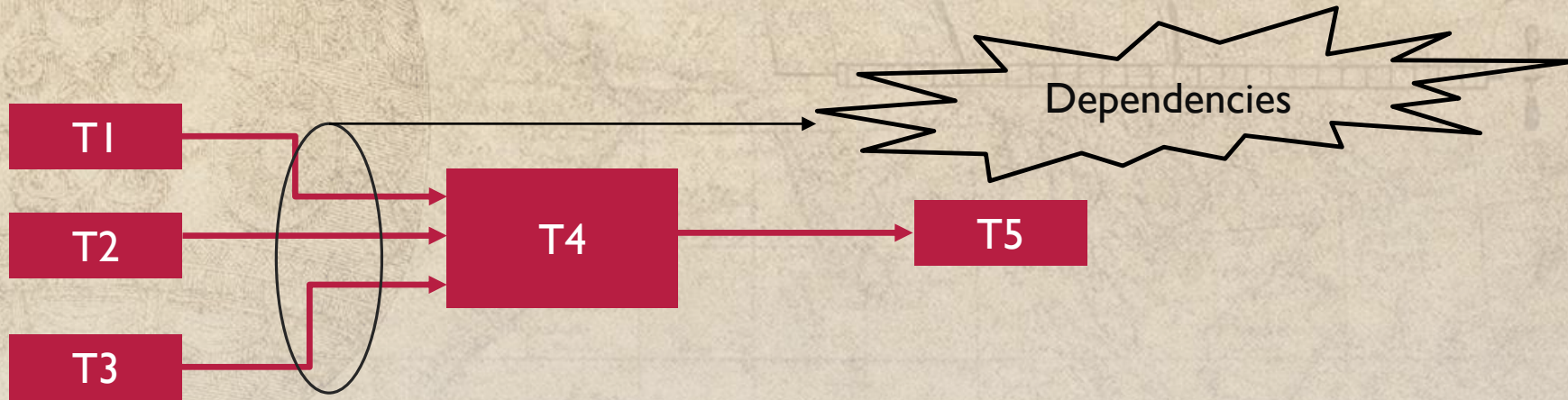- *Note : Workflows are defined in python, but the steps themselves can be in bash, python function, DB function etc*

What is Apache airflow / Cloud composer?

- It helps to automate scripts in order to perform tasks.

T1 → T2 → T3 → T4 → T5

- If this is the case, do we require apache airflow since it in sequence? But, what if each task takes 2 hours time, which mean totally it takes 10 hours to complete this workflow for a meaningful results. However the business expectation is to complete in 6 hours?.

- Lets, assume we have identified T1, T2, T3 can start simultaneously and once they are completed successfully then T4 can start.

What is Apache airflow / Cloud composer?

T1

T2

T4

T5

T3

Dependencies

- T1, T2, T3 are triggered to start at same point of time and once all of them are complete then T4 can start.

- A task belongs to a DAG - it is where you specify the operations to perform