BIG DATA

4 - Le cycle de vie de la donnée

SOMMAIRE

ORDONNANCEMENT

DÉFINITION

L'ordonnancement permet de définir des enchaînements entre les traitements.[...]
Les systèmes évolués permettent des enchaînements conditionnés suivant les types de fin (normale, anormale, spécifique ou quelconque) de traitement(s) précédent(s), la détection et / ou le contenu de fichier(s), l'arrivée de mail(s) avec ou sans pièce(s) jointe(s), un traitement sur une base de données, …

Wikipédia

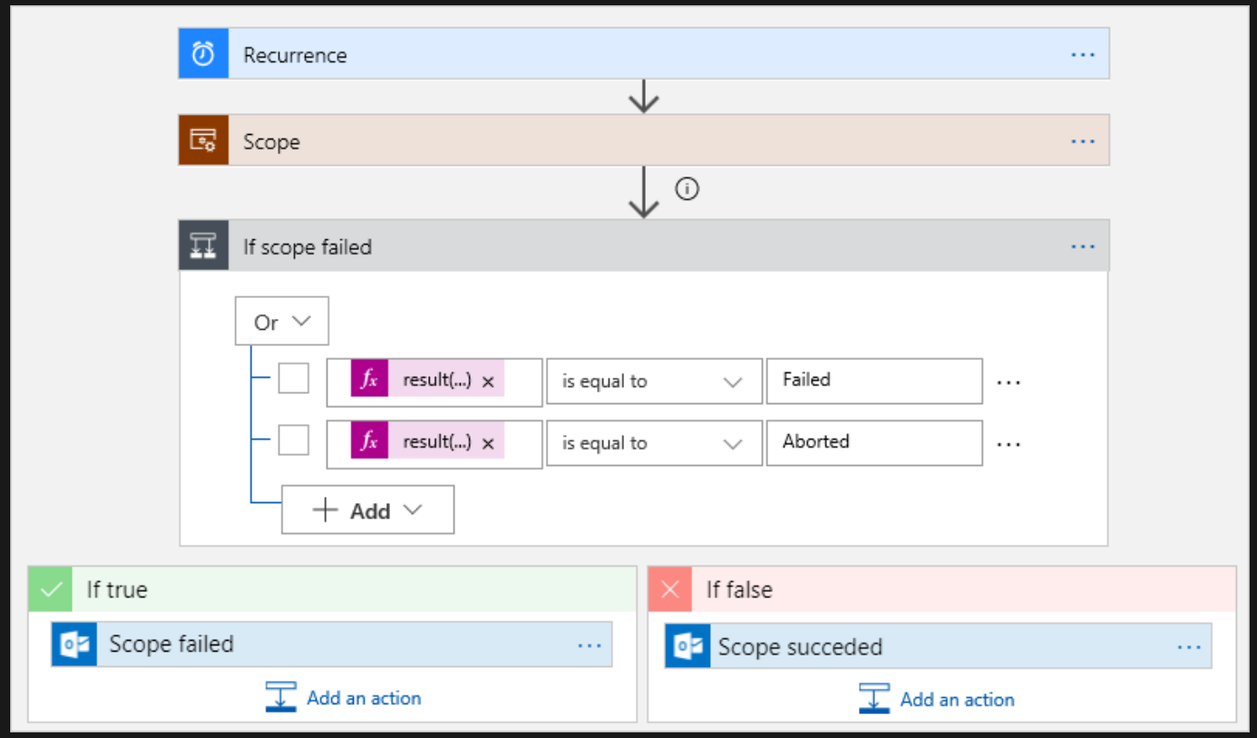
ORDONNANCEMENT

OBJECTIFS

- Automatisation : planification calendaire, déclenchement sur évènements
- Distribution des tâches : différents outils, différentes ressources
- Scénarios complexes : dépendances, enchainements, conditions, etc.
- Supervision : Récupération de codes erreurs, tâches correctives

ORDONNANCEMENT

EXEMPLE (AZURE LOGIC APPS)



ORDONNANCEMENT

LES OUTILS

- Hadoop : Oozie
- Azure : Logic Apps
- AWS : Batch
- GCP : Cloud Scheduler

L'INGESTION DE DONNÉES

DÉFINITION

L'ingestion des données désigne l'étape du processus de préparation des données pendant laquelle les données provenant d'une ou plusieurs sources vont être placées dans un espace de stockage.
Le but est de permettre aux équipes d'avoir accès à ces données pour pouvoir les utiliser, les organiser ou encore les analyser.

Wikipédia

L'INGESTION DE DONNÉES

OBJECTIFS

- Collecte : récupération des données de systèmes connexes
- Centralisation : mise à disposition à un emplacement unique
- Isolation : diminuer l'adhérence aux autres systèmes
- Historisation : rejeu

L'INGESTION DE DONNÉES

LES OUTILS

- Hadoop : CLI, NiFi, Kafka
- Azure : CLI, Data Factory
- AWS : CLI, Glue
- GCP : CLI, Dataflow

LA PRÉPARATION DE LA DONNÉE

DÉFINITION

La préparation de données est un processus qui précède celui de l'analyse de données.
Il est constitué de plusieurs tâches comme la collecte de données, le nettoyage de données,
l'enrichissement de données ou encore la fusion de données.

Wikipédia

LA PRÉPARATION DE LA DONNÉE

OBJECTIFS

- Exploitabilité : données brutes => données exploitables et structurées
- Nettoyage : uniformisation, suppression, etc.
- Enrichissement : opendata, référentiels internes, etc.
- Fusion : agrégation de plusieurs référentiels

LA PRÉPARATION DE LA DONNÉE

LES OUTILS / LANGUAGES

- Python, PySpark, R, Scala
- SQL
- Talend, Informatica, Alteryx, etc.
- etc.

L'EXPOSITION

DÉFINITION

A l'inverse de l'ingestion, l'exposition consiste à mettre à disposition de la donnée pour les autres systèmes.

L'EXPOSITION

OBJECTIFS

- Données structurées et exploitables
- Référentiels
- KPI
- ML/IA : modèles entrainées
- DataViz : graphiques, cartes, nuage de mots, etc.

L'EXPOSITION

LES OUTILS

- Hadoop : Hive, Hue
- Azure : Synapse, Power BI
- AWS : RedShift, QuickSight
- GCP : BigQuery, Data Studio

QUESTIONS ?