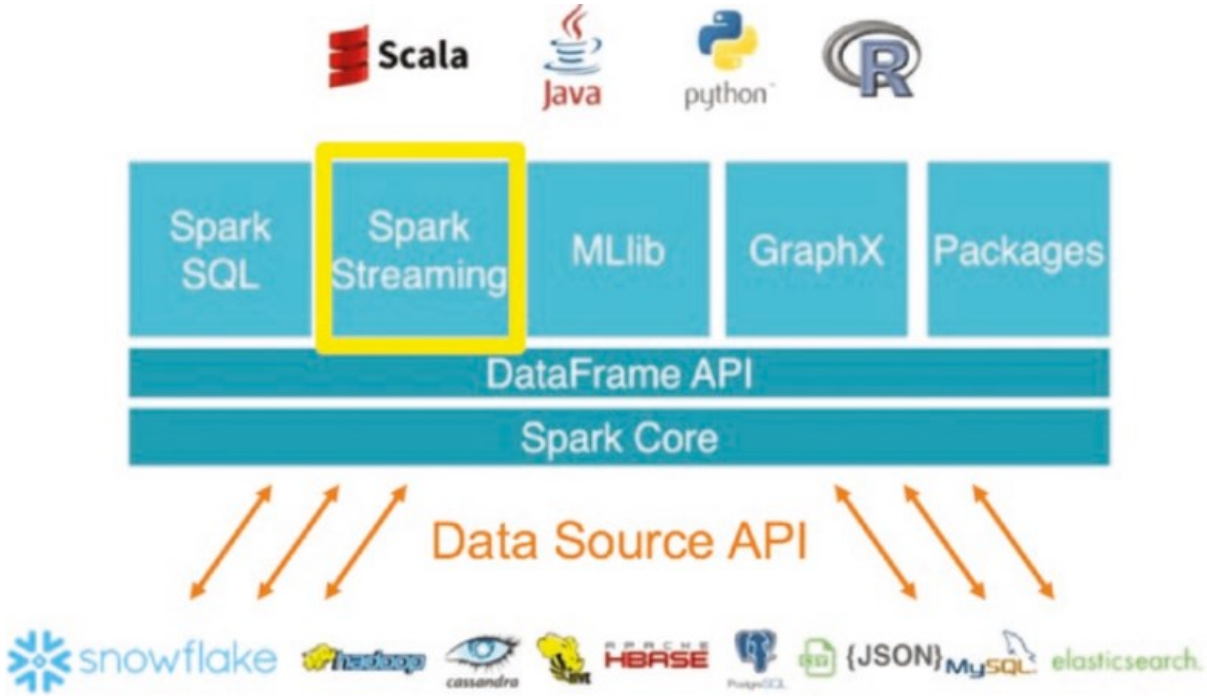
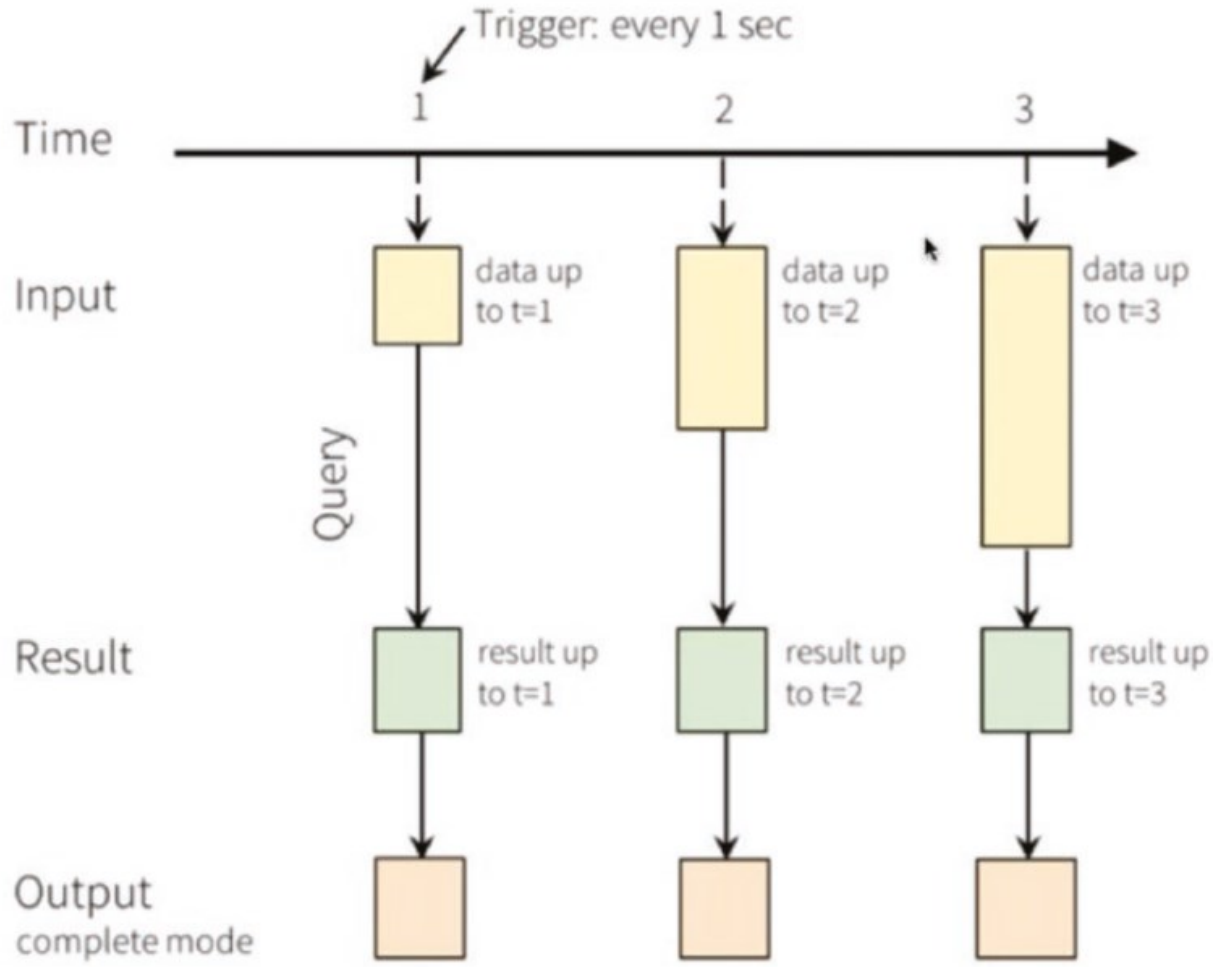
POINTS IMPORTANTS DU CHAPITRE 3 :SPARK STRUCTURED STREAMING

- La différence entre traitement de données par lots et par flux est que :
  - le lot fait référence à un groupe d'enregistrements rassemblés sur une période de temps et utilisés ultérieurement pour le traitement et l'analyse. Étant donné que ces enregistrements sont collectés sur une période de temps, en termes de taille, les données de lot sont généralement plus volumineuses que les données de flux (dans certains cas, cependant, les données de flux peuvent être plus volumineuses que les données de lot) et sont souvent utilisées pour effectuer des post-mortem à diverses fins d'analyse.
  - Le traitement de flux fait référence au traitement des enregistrements en temps réel ou quasi réel. On n'attend pas la fin de la journée pour ensuite traiter ou analyser les données. Au lieu de cela, les enregistrements de l'ensemble de données sont traités un par un dès qu'ils deviennent disponibles ou sur la base d'une période de fenêtre.
  - Les entreprises veulent utiliser les données les plus récentes ou les plus récentes pour générer des informations utiles qui peuvent aider à la prise de décision. Le traitement par lots ne peut pas offrir d'analyse à la volée, car il ne fonctionne pas en temps réel, tandis que le traitement des données en continu peut aider plus efficacement dans des cas tels que la détection de fraude.
- Spark Streaming l'un des composants du framework Spark presente cette architecture :



- Structured Streaming : est la dernière version du composant de streaming dans Spark et offre la même API pour les travaux de traitement de données par lots et par flux. Son processus est la suivante :



- Les trois domaines principaux dans lesquels nous pouvons diviser ce cadre de diffusion de streaming structuré :
  - Data input
  - Data processing (real time or near real time)
  - Final output
-