Streaming /Realtime

24 /7 - continuous

Chronologically Order is very important

Size of the data → kb

Tools :

Data ingestion tools /

data storage/data processing

Kafka --streaming platform

Spark streaming /structured streaming

Flink

Storm

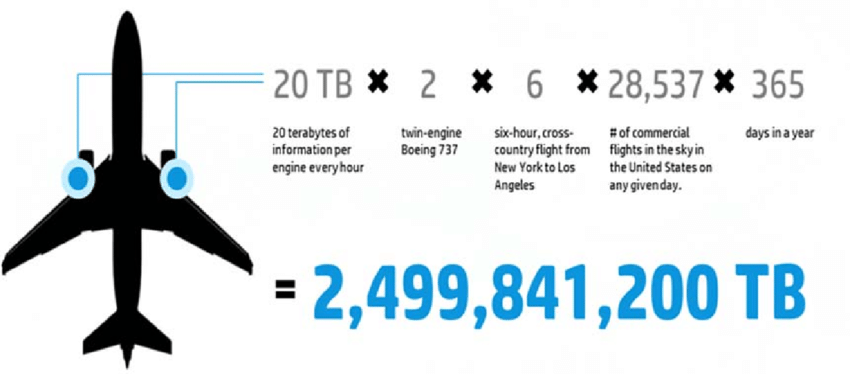
Flume

AWS kinesis

Azure event grid/eventhub

===================================

Use cases of streaming data in different domains



=========================================

Spark components

Pyspark

Spark-shell

REPL ⇒ Read Evaluate Print Loop

Spark → fundamental unit → RDD

RDD → Resilient Distributed Datasets

RDD → collection of partitions

Features of RDD:

1.Immutable

2.type inferred

3. Cacheable

4. Lazy evaluation

5.fault tolerance

Spark operations :

Transformation → creates new RDDs

and Action → Triggers the job and gives the output