

Map Reduce and Apache Spark do the same thing. But Spark is 100 times better than Map Reduce.

PIG is high level environment to write Map Reduce.

Flume is to transfer data to and from HDFS.

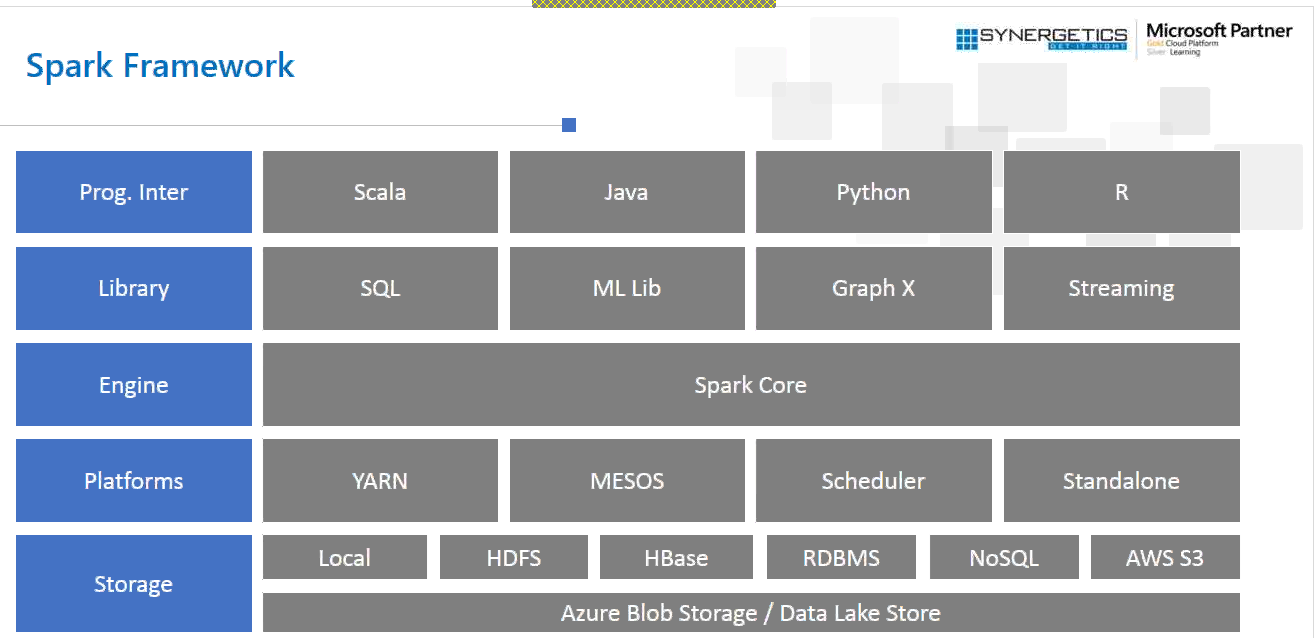
KAFKA: Distributing messaging system.





Map reduce is very slow.

Apache storm is for streaming only.



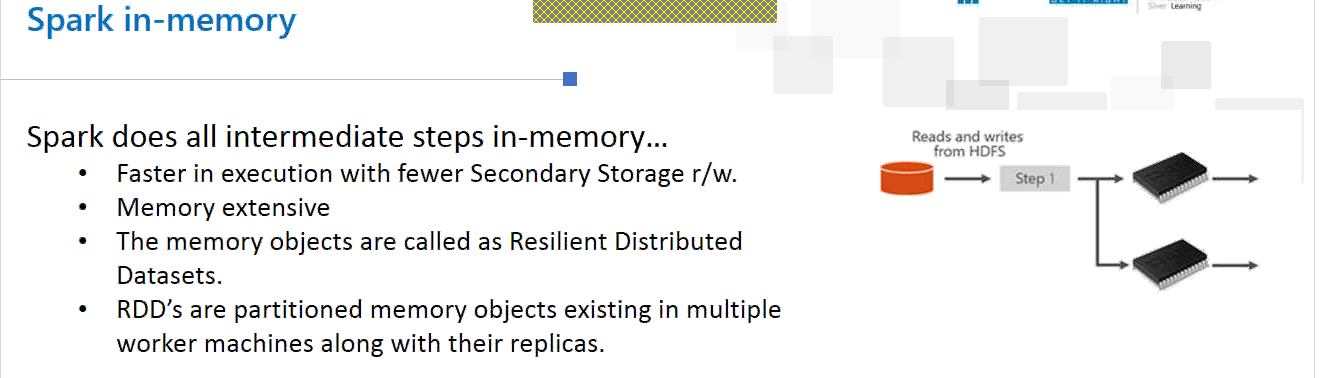
Spark is written in Scala and java.

R has extensive library of analytics and Machine Learning.

Spark stores every data in memory and not to disk.

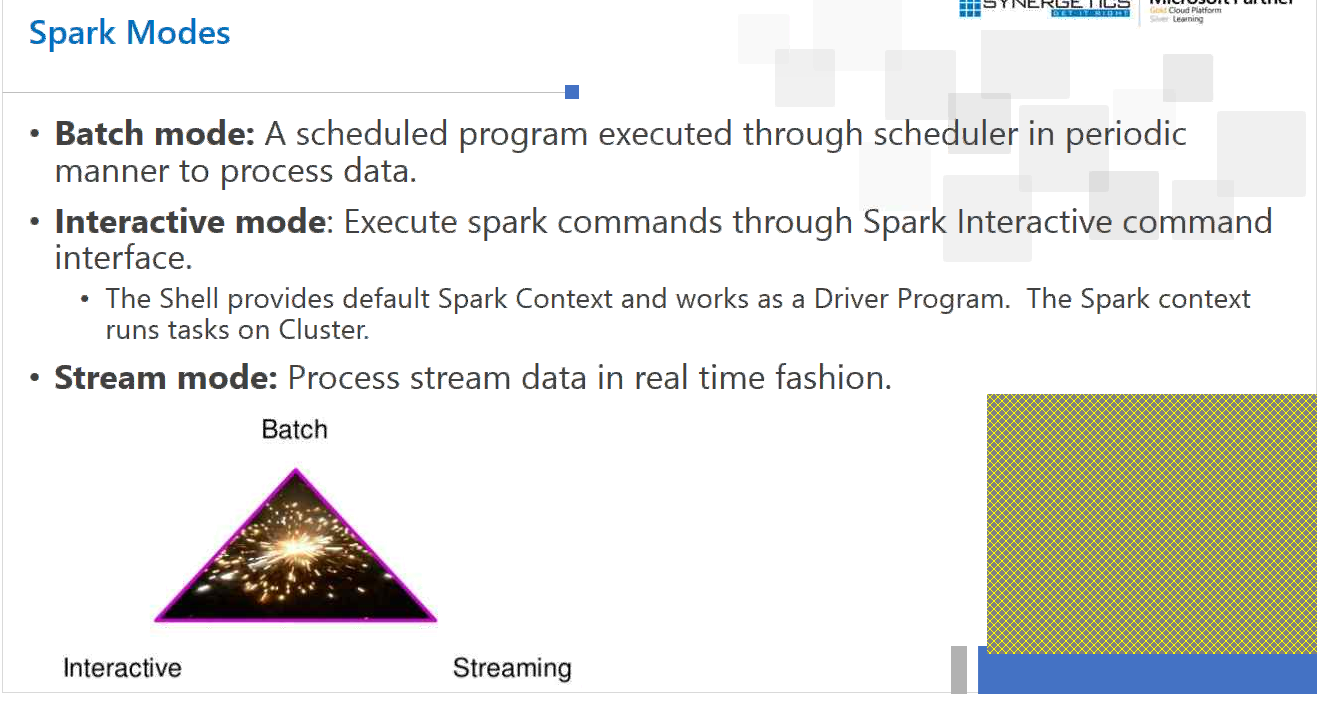


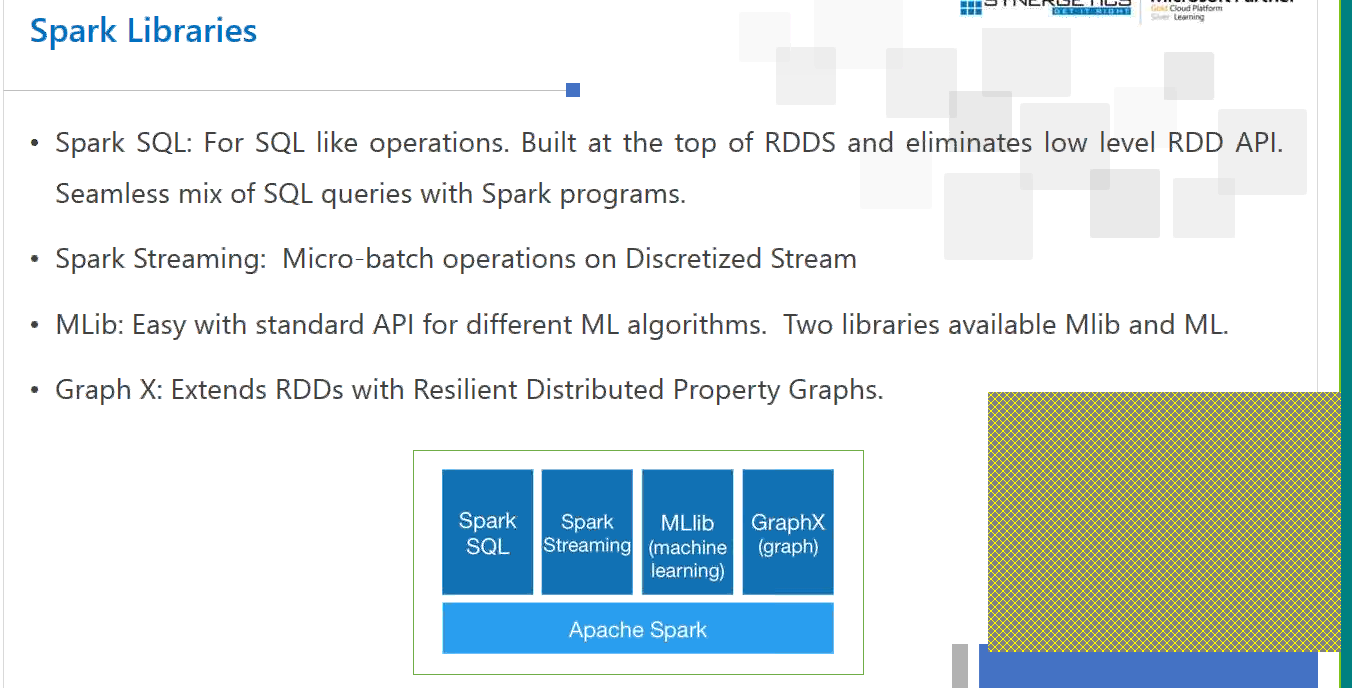
As data is not stored in hard disk in spark rather it keeps in memory. So processing is faster in spark.

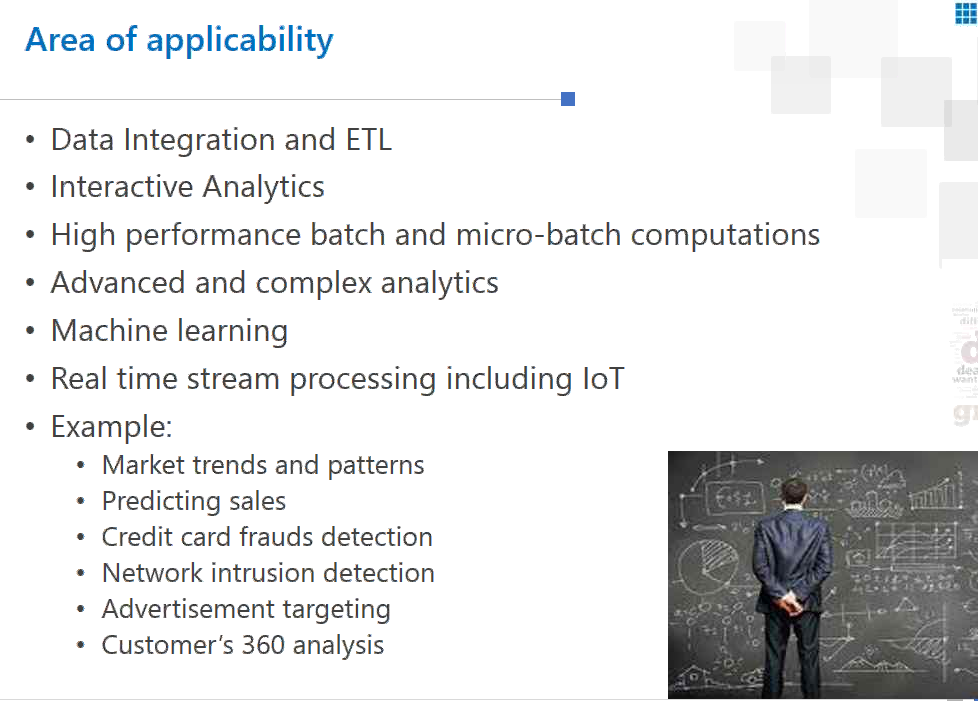


RDD is collection of object where each object may belong to different machine.

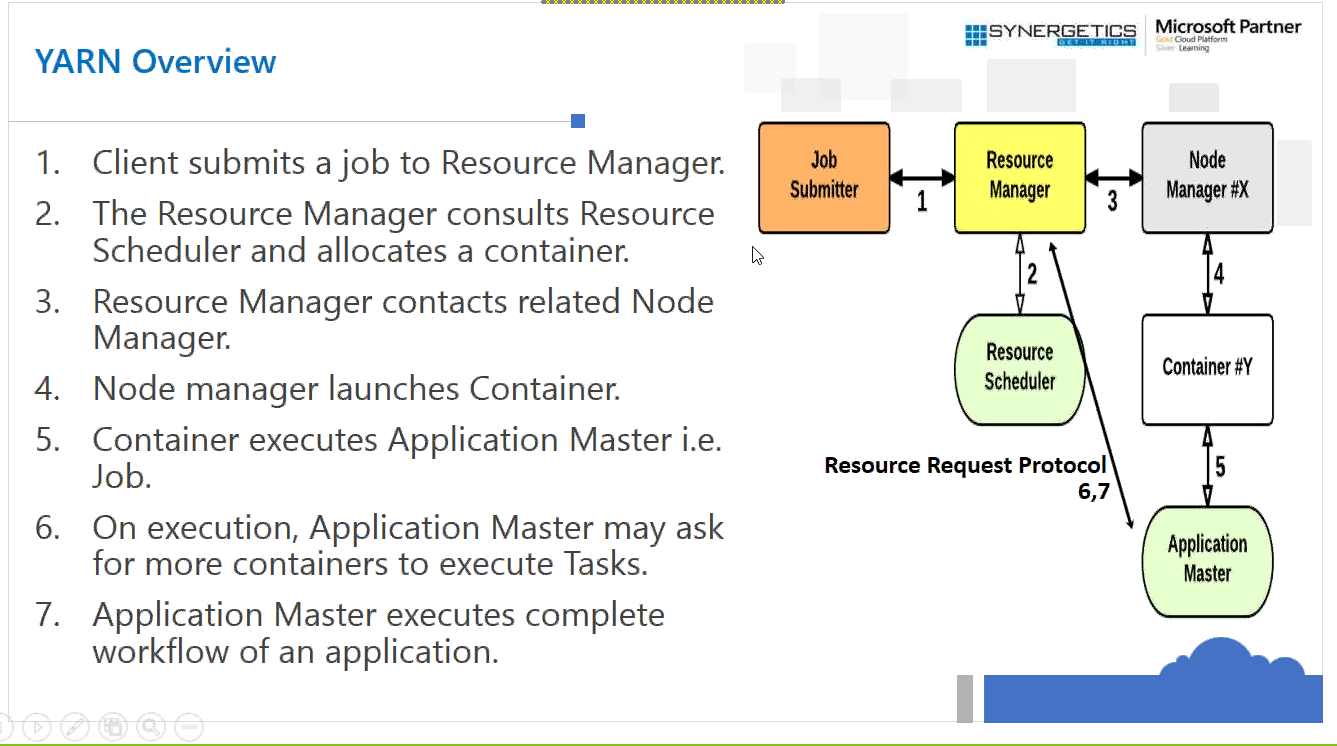
YARN🡪Resource Master and Resource Scheduler. The Scheduler is used to handle the machine in Spark.

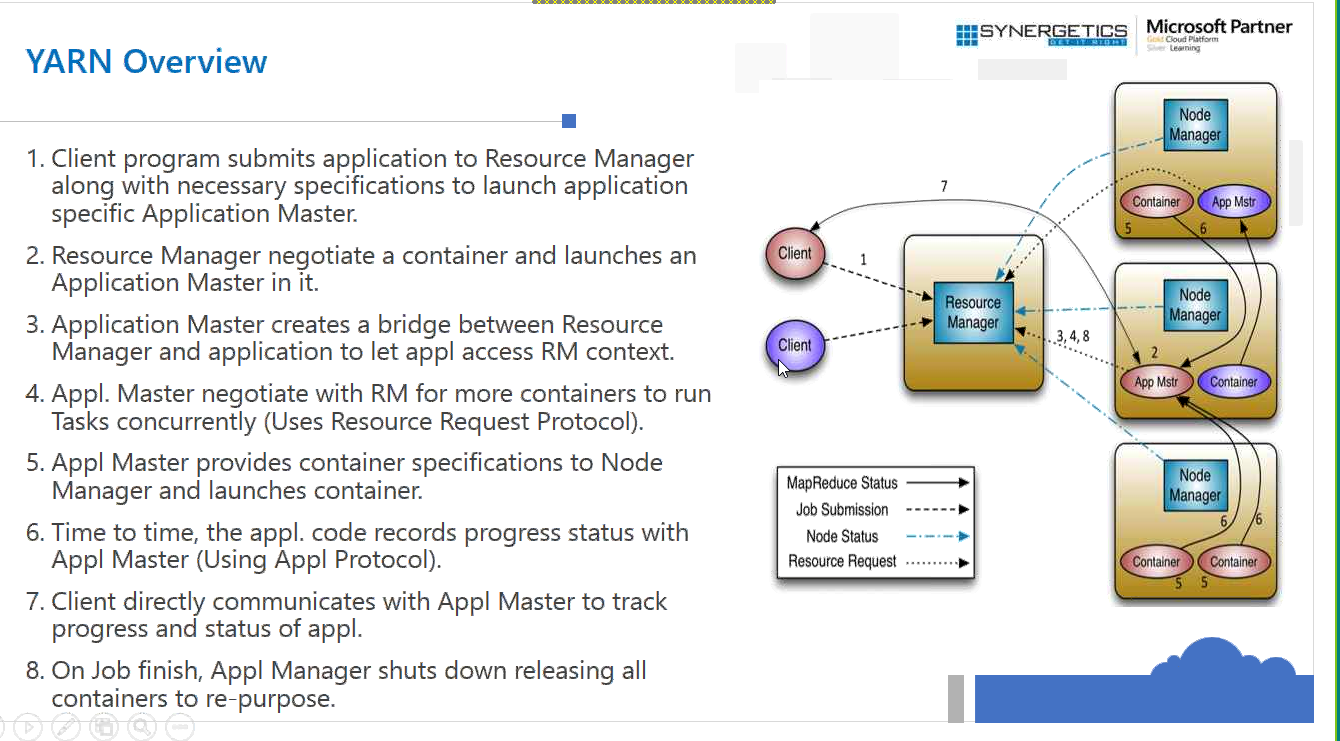




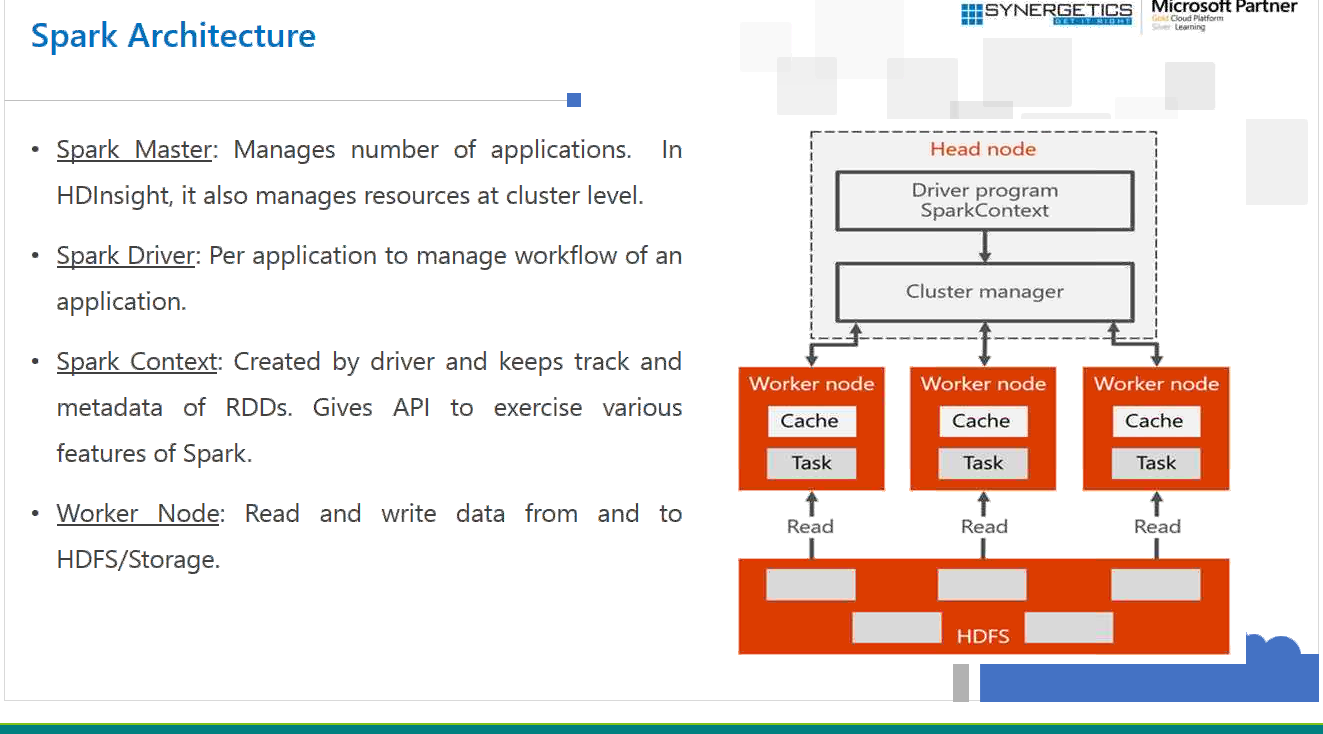


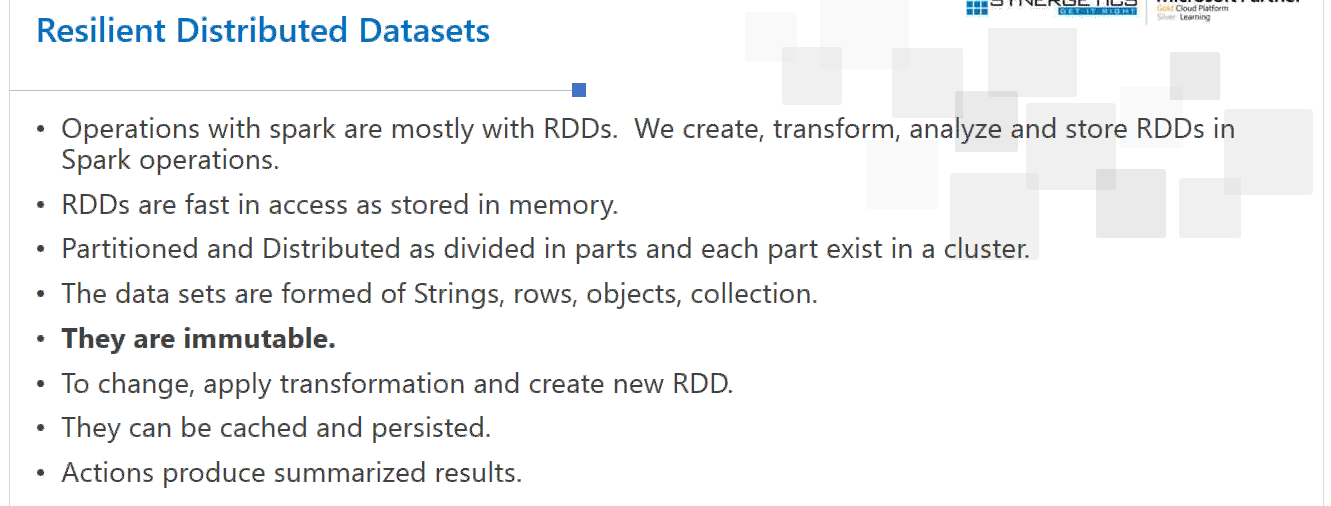
YARN:

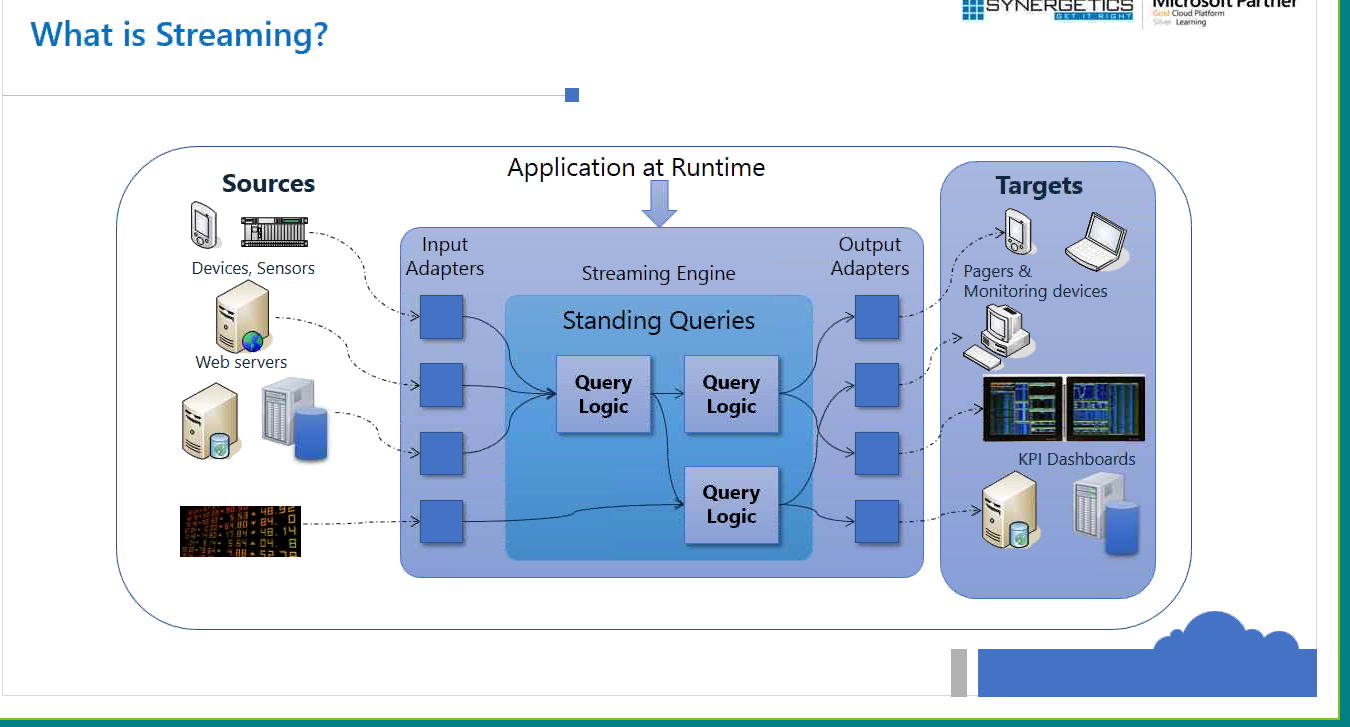


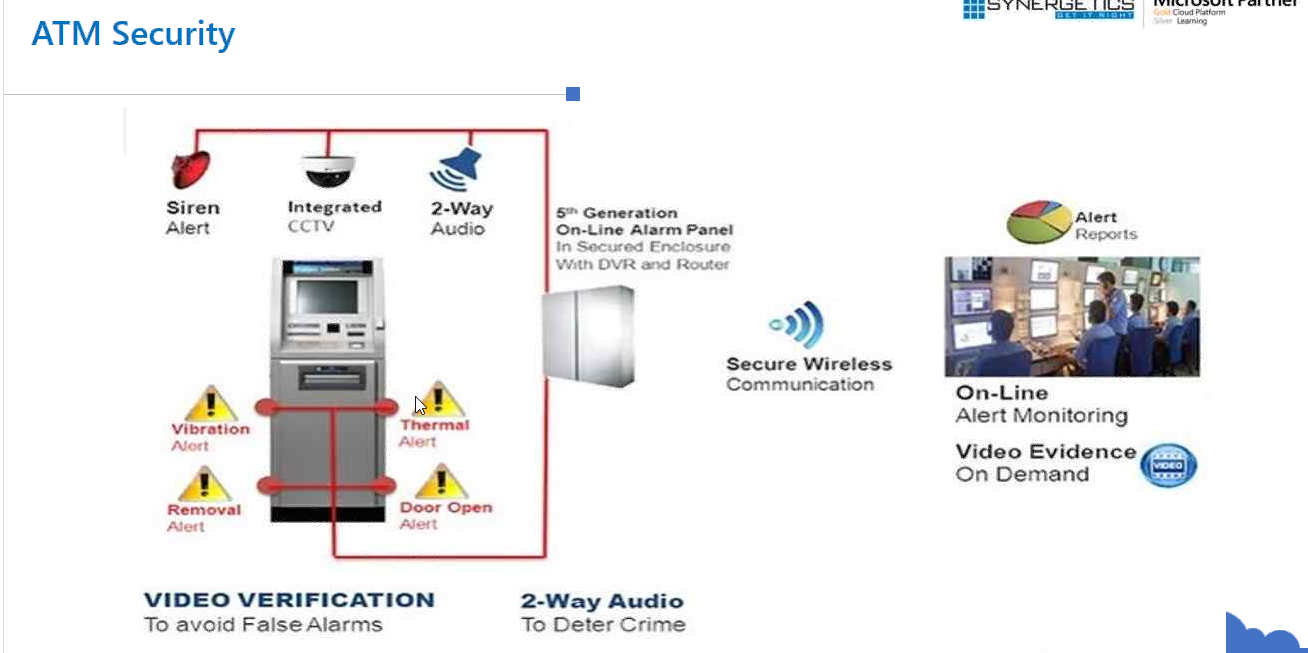


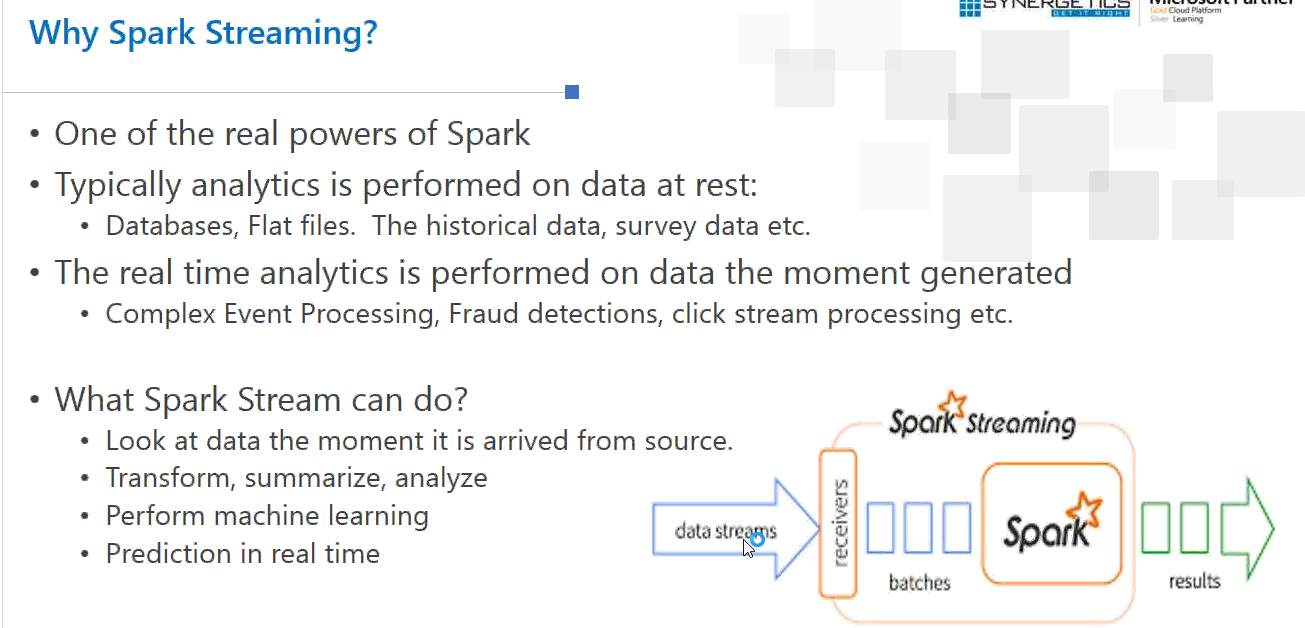
One application==One Application Master. One Job ==One Master

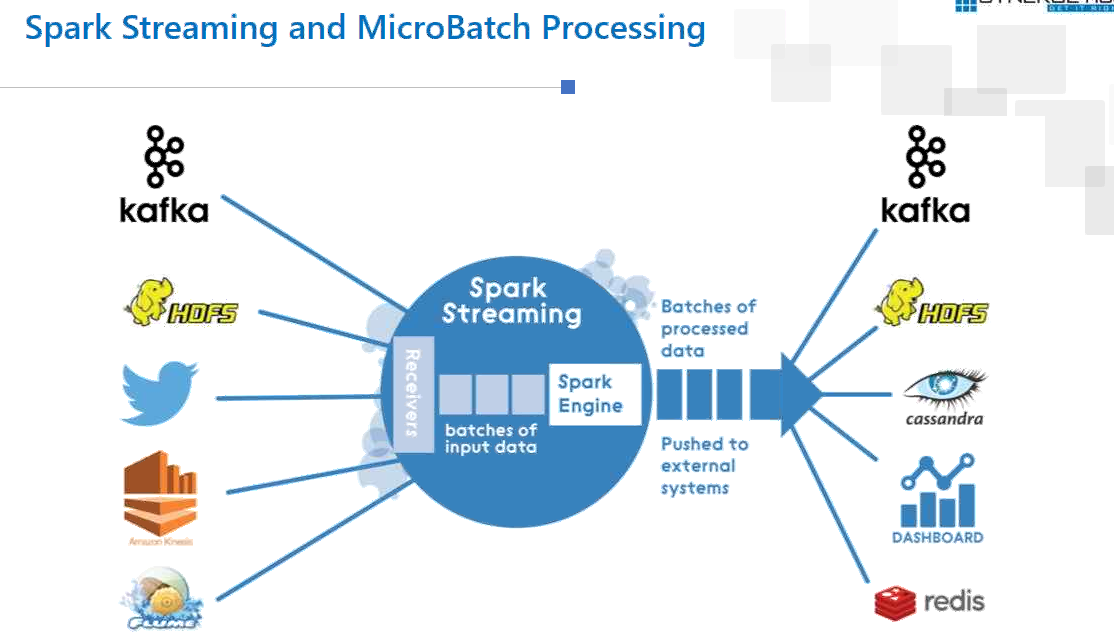




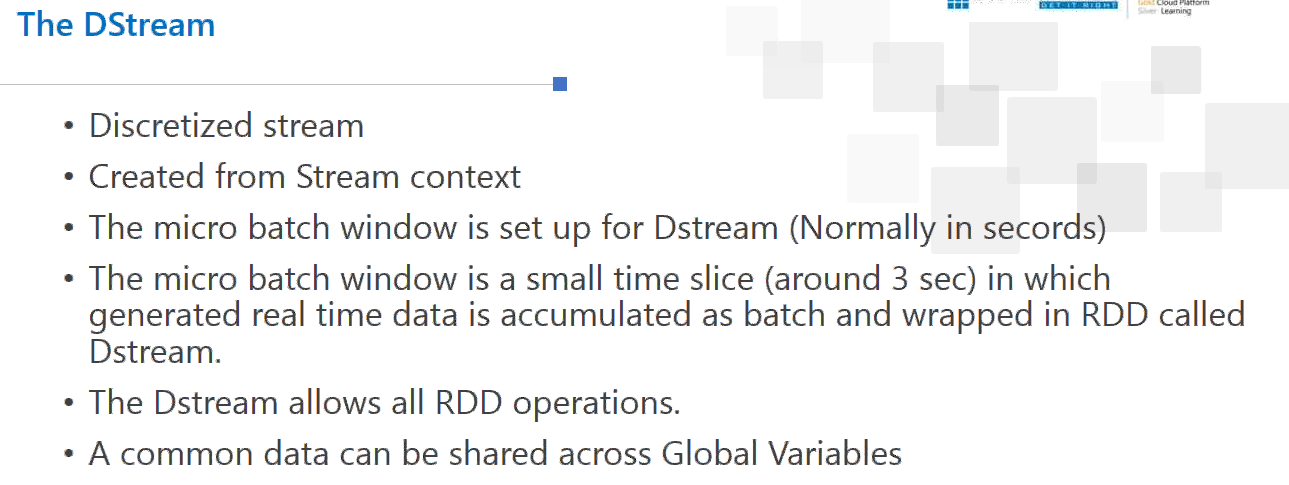








KAFKA = High speed messaging system.



PySpark

