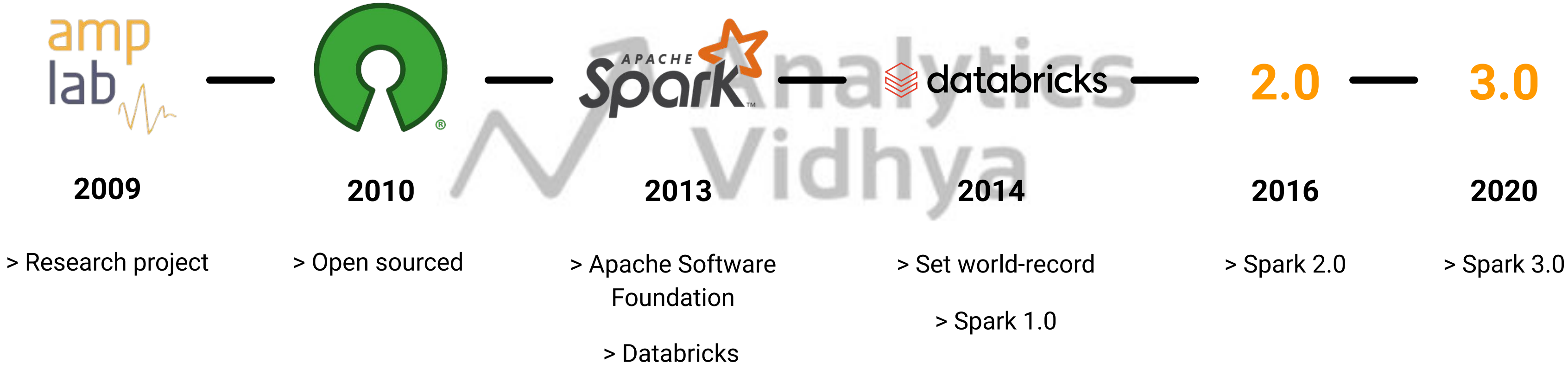
Introduction to Apache Spark

What is Apache Spark?

Apache Spark is a parallel data processing engine for big data and machine learning applications designed to run on a cluster of computers.

History of Apache Spark



2009

> Research project

2010

> Open sourced

2013

> Apache Software Foundation

> Databricks

2014

> Set world-record

> Spark 1.0

2016

> Spark 2.0

2020

> Spark 3.0

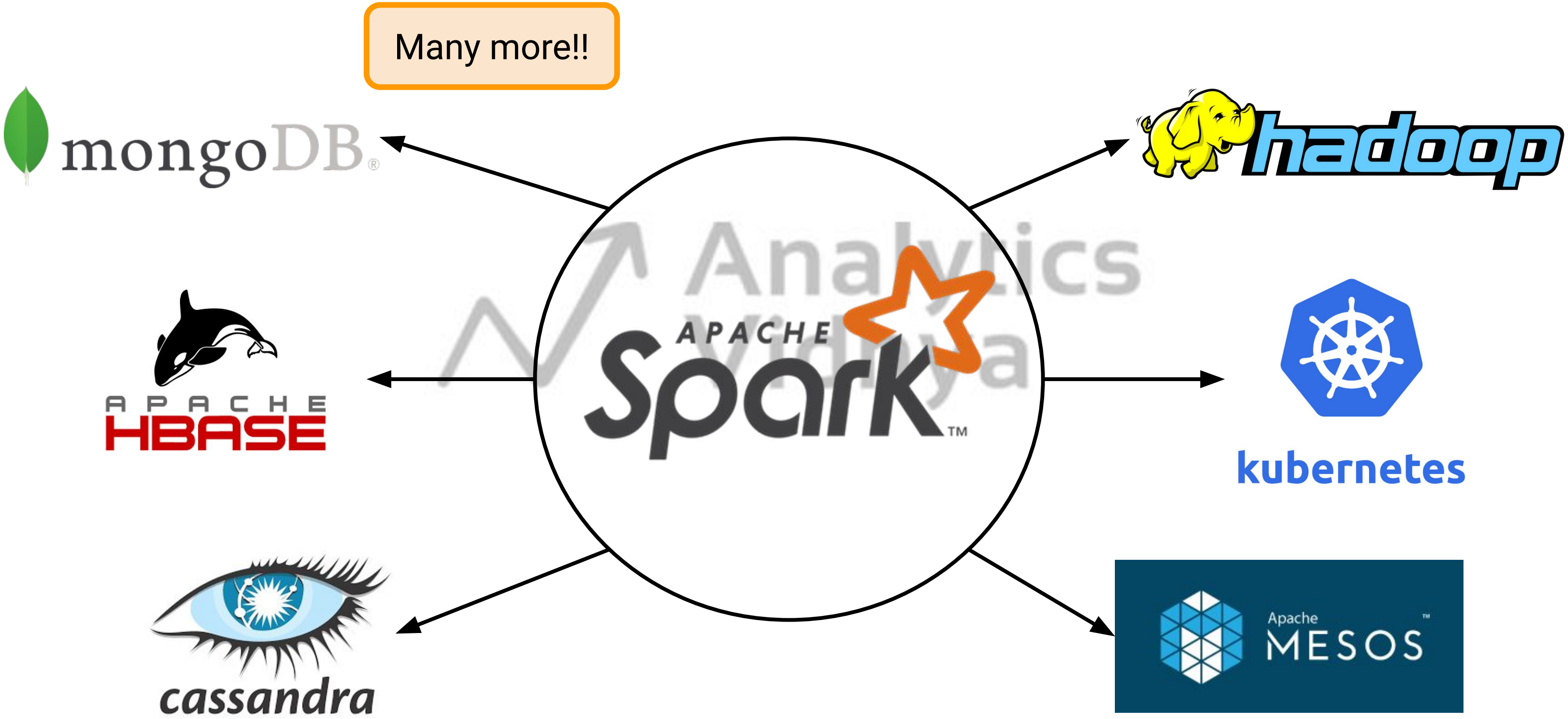
Features of Apache Spark

- Polyglot
- Flexibility
- Unified Engine
- In-memory computation
- Real-time Stream Processing

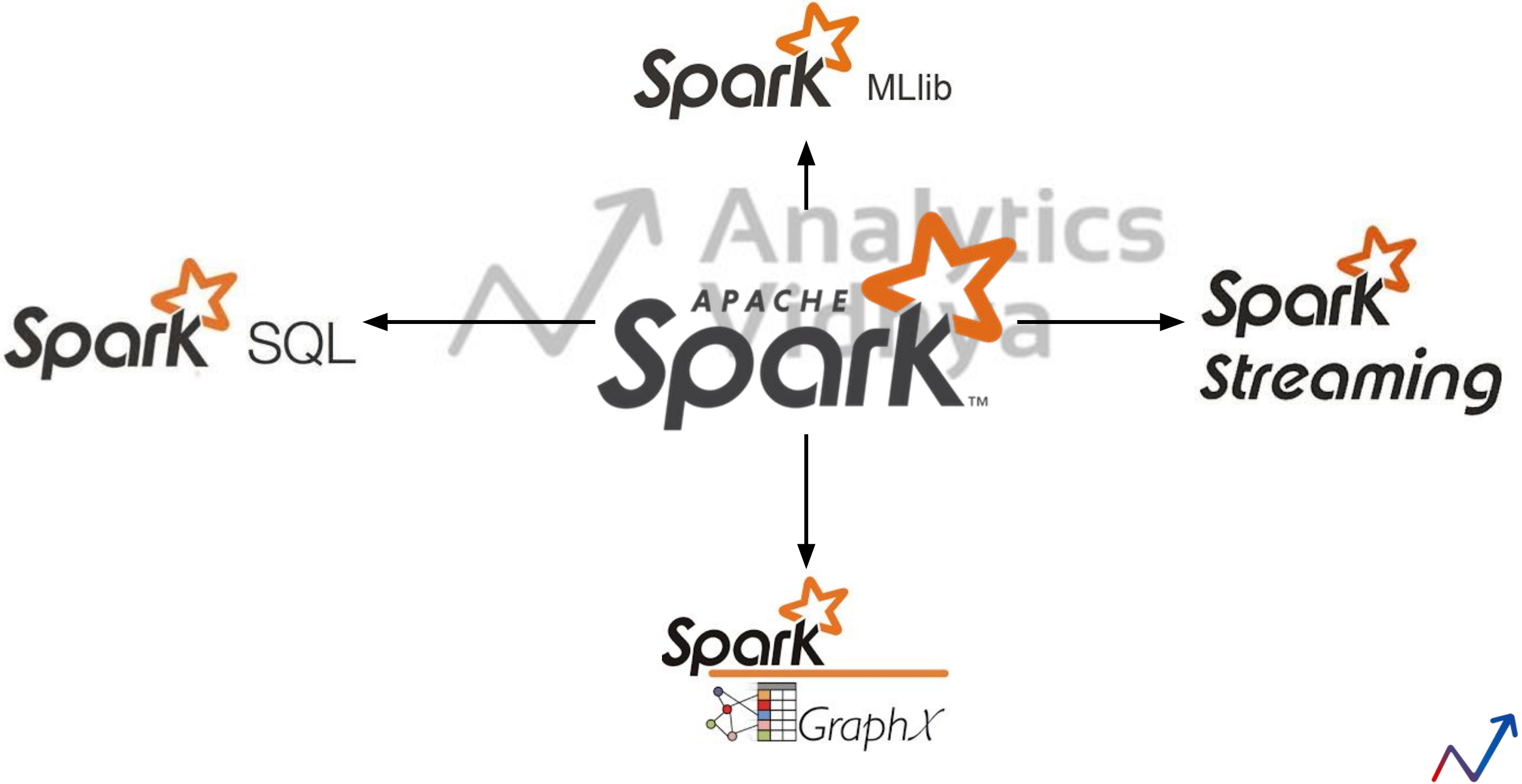
Polyglot

Java
R
APACHE
Spark
Scala

Flexibility

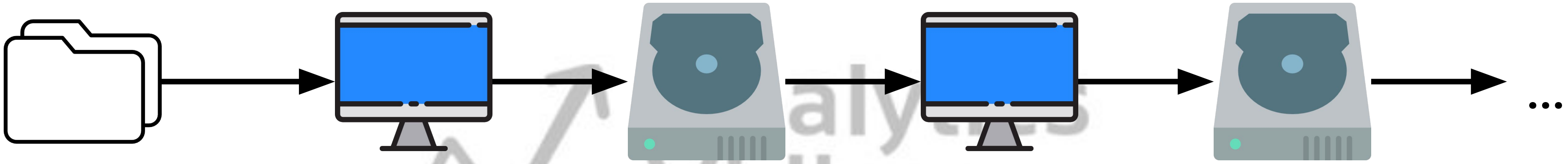

Many more!!

Unified Engine



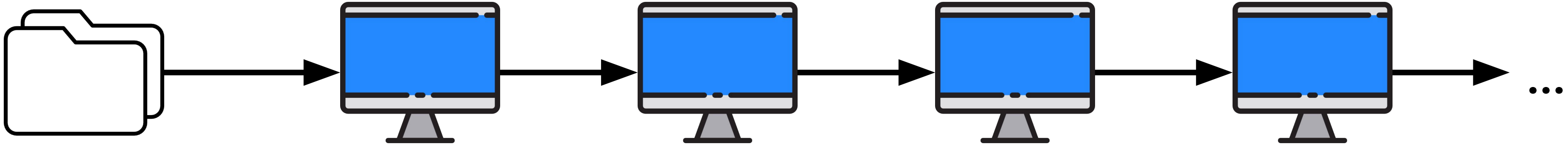
In-memory computation

Hadoop MapReduce

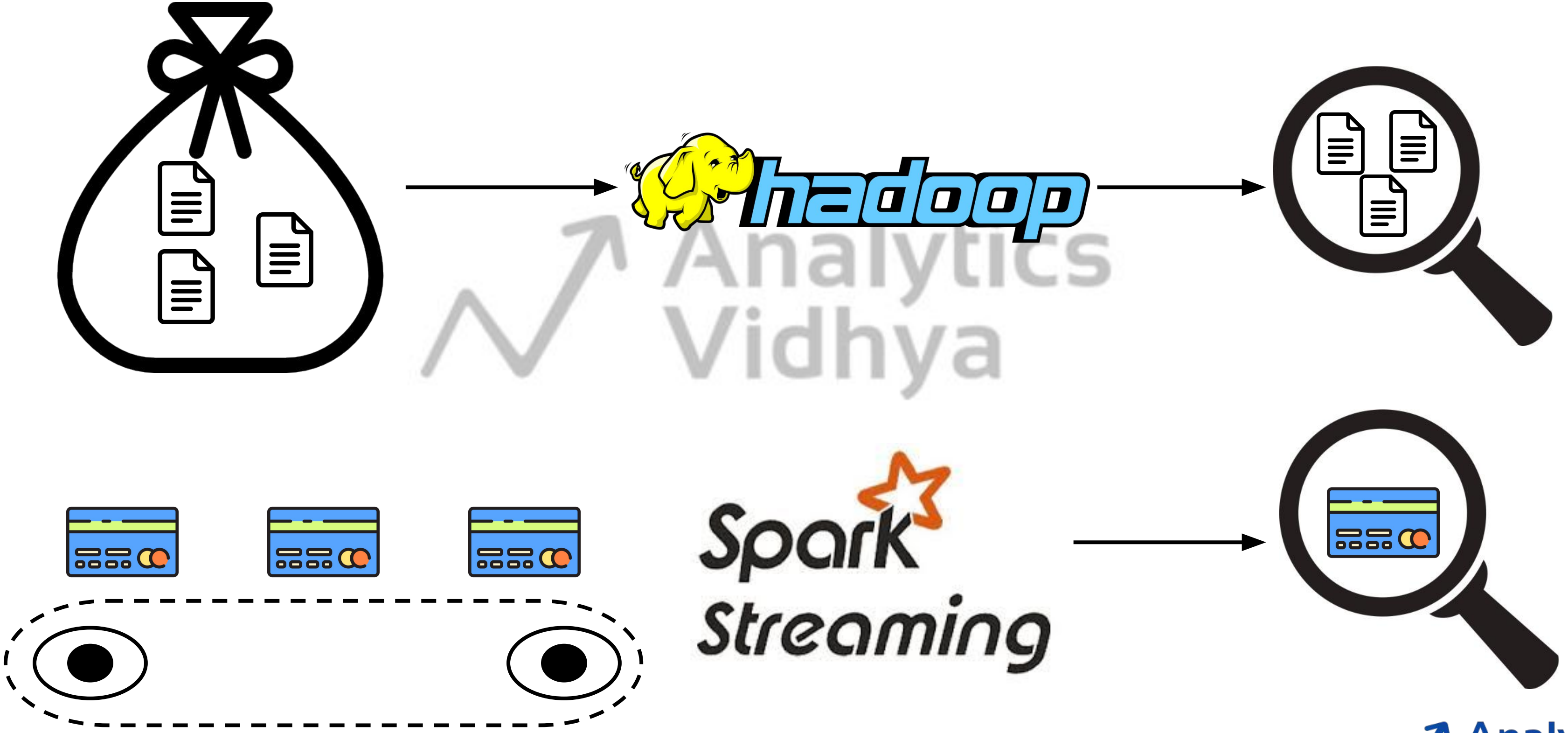


Apache Spark

100x faster!!



Real-time stream processing



Use cases of Apache Spark

Use cases of Apache Spark

- 103 millions monthly ride hailers in over 900 cities worldwide
- 100, 000+ Spark applications run everyday
- Diverse data sources: HDFS, Hive, Cassandra, MySQL, and more

Use cases of Apache Spark

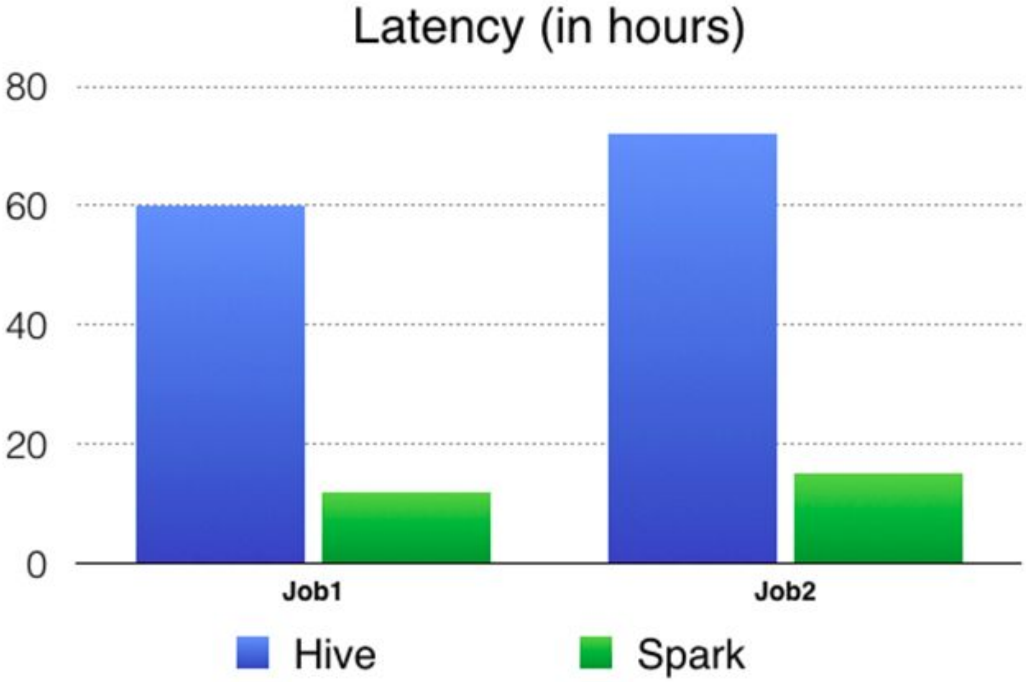
- 200 million paid-subscribers
- Content personalization
- Spark streaming for personalized videos on homepage

Use cases of Apache Spark

facebook

- More than 2 billion monthly users
- Hive for multiple analytics tasks
- Hive vs Spark

60 TB data handled with Hive and Spark



Thank you!