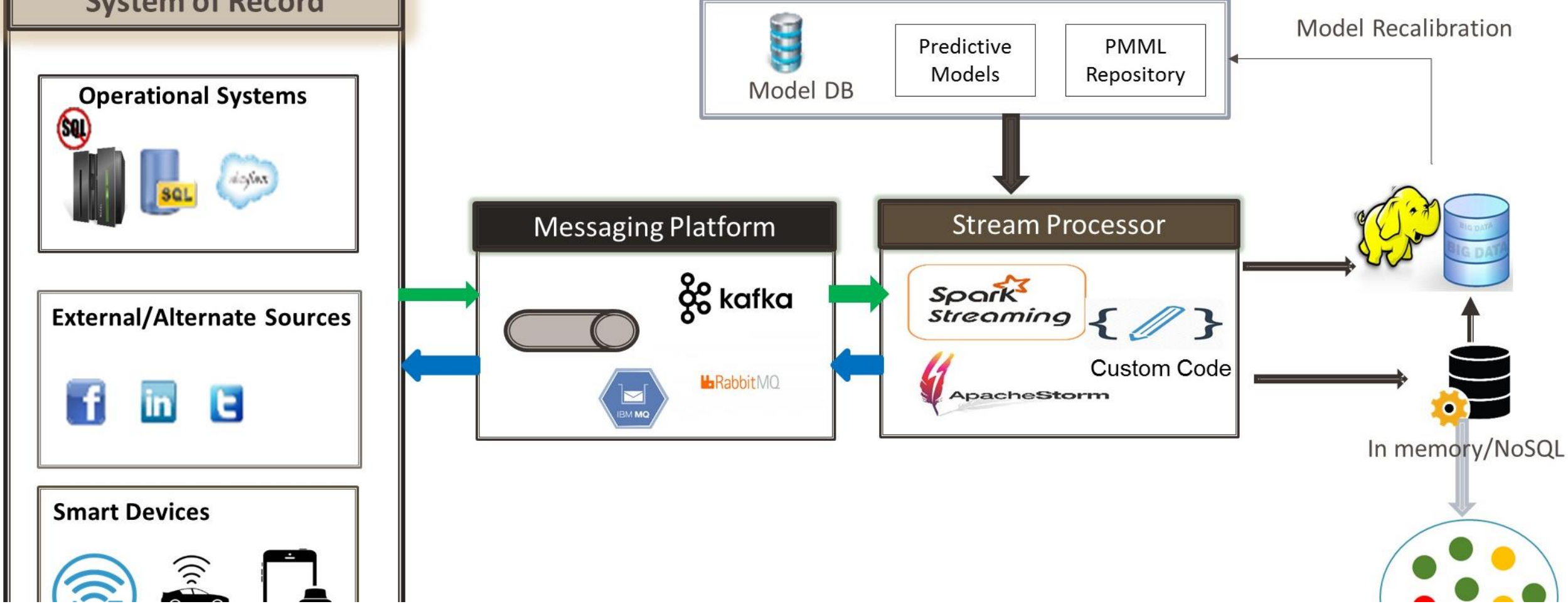
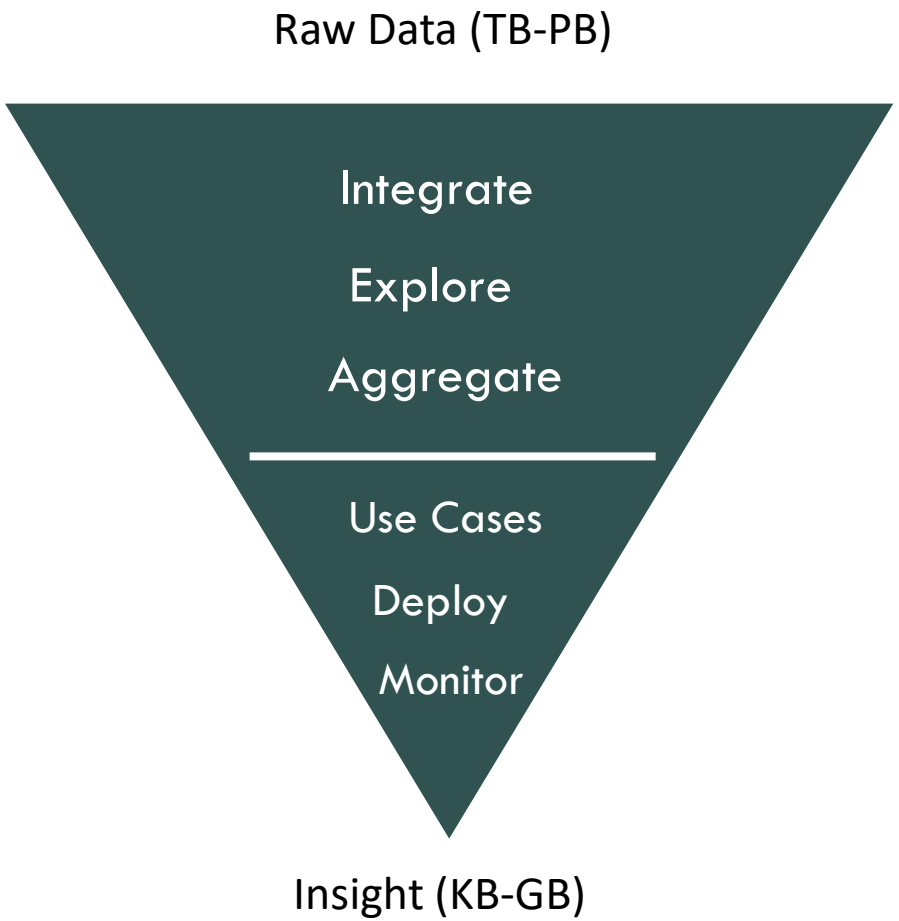
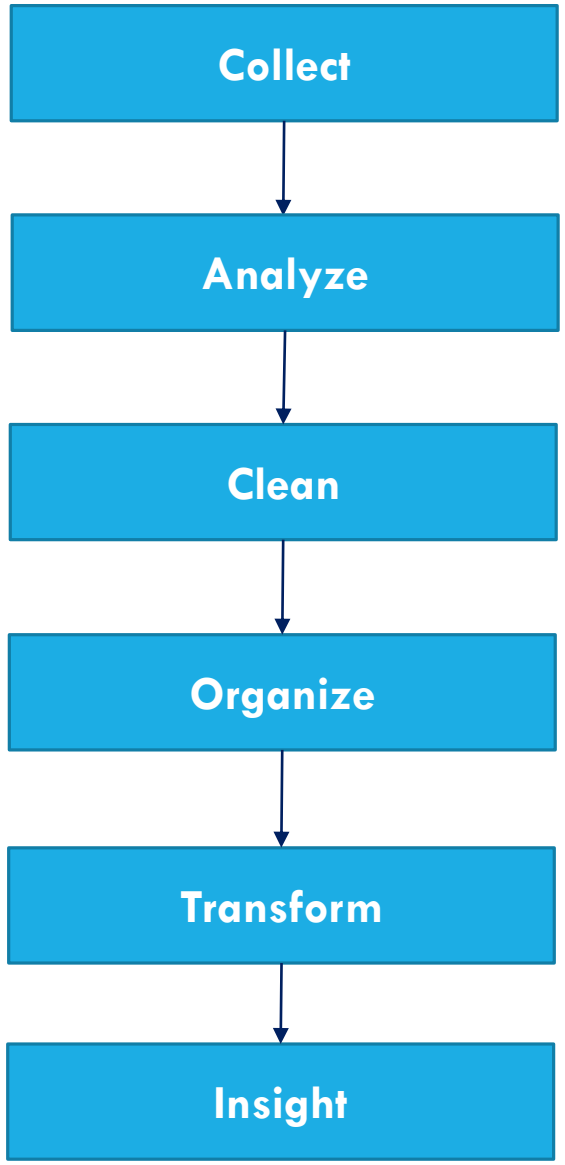


SPARK FOR DATA ENGINEERING

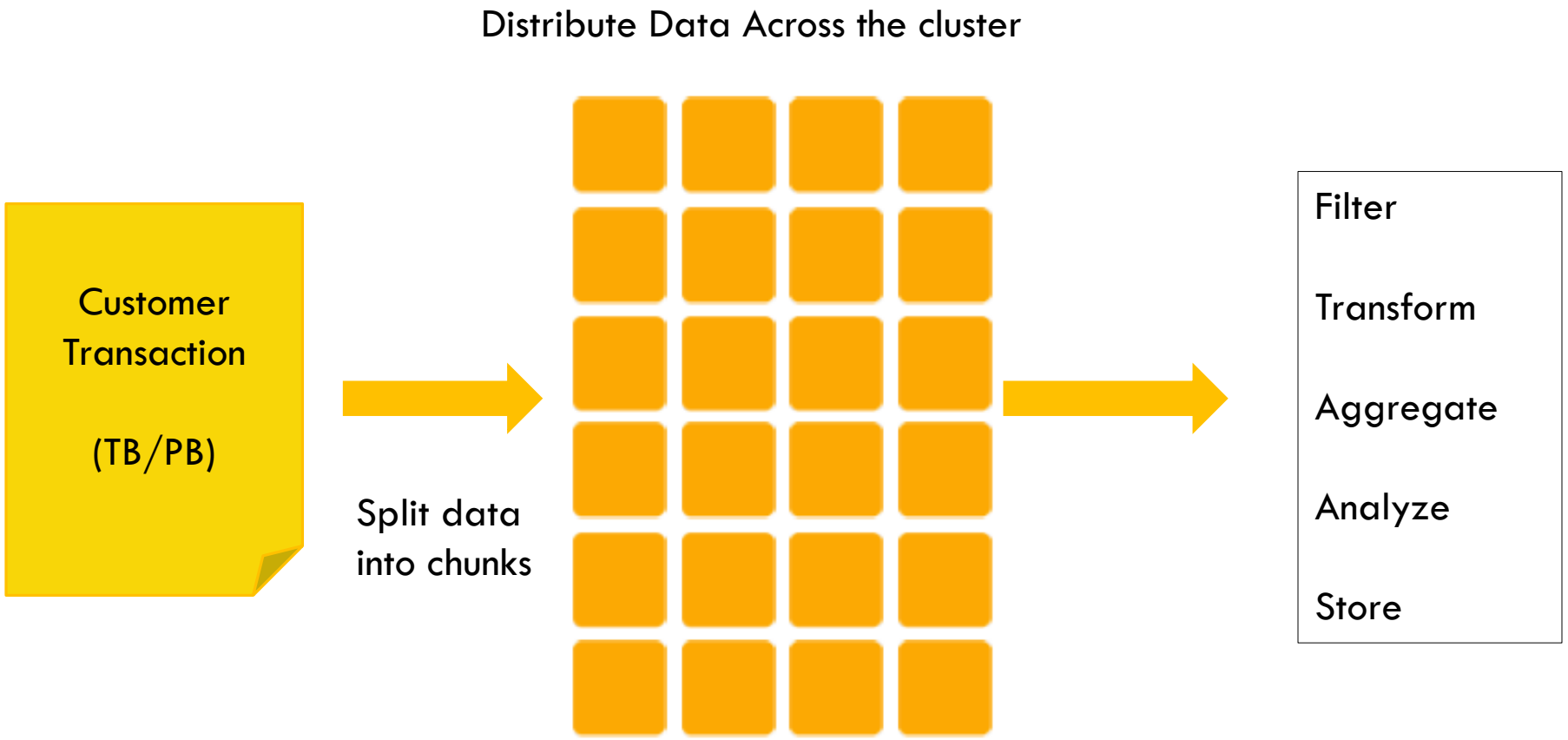
Background

DATA LIFECYCLE





HOW SPARK WORKS?



WHY SPARK?

Data Collection

Batch

Real Time

Data Processing and Transformation

Apache Spark

Spark Streaming

Storage and Analysis

Amazon S3

Google Cloud Storage

Microsoft Azure Blob Storage

cassandra

APACHE HBASE

WHY SPARK?

Amazon S3, Google GCS, Azure, HDFS

Kafka, Kinesis, pub sub

Baremetal

Cloud Managed Services

Cloud Kubernetes Services

WHY SPARK?

Baremetal	Cloud Managed Services	Kubernetes Services
Yarn	 	   
Standalone		
Apache Mesos		
Kubernetes		

SPARK ARCHITECTURE