Web: Inceptez.com Mail: info@inceptez.com Call: 7871299810, 7871299817

# APACHE SPARK


# INTRODUCTION

# COURSE HIGHLIGHTS

- Extensive, In-depth & Comprehensive
- Use cases and Hands on based
- Designed as per market requirements that covers
  - Batch Processing
  - Real-time Streaming
  - In-memory processing
  - RDBMS Datastore
  - Columnar & Document Datastore
  - Visualization and Dashboard
- Latest Versions
  - Spark 2.2.0 - Scala 2.11.1 - Java 1.8.0 - Hadoop 2.7.1 - Hive 1.2.2 - Cassandra 3.9.0 - ElasticSearch 5.0.1 - Kibana 5.0.1 - Kafka 0.10 – Nifi 1.5.0 - Centos 7
- End to End System Integration
  - Acquire -> Transport -> Queue -> Transform -> Enrich -> Lookup -> Stream -> Load -> Visualize
- Performance Tuning

# COURSE AGENDA

- BIG DATA Overview
- Hadoop Architecture
- Spark Basics
- Scala and its programming implementation
- Spark Execution Model
- Working with RDDs
- Spark Essentials
- Spark Distribution Setup and Configurations
- Running Spark on Cluster
- Writing Spark Applications
- RDD Operations in detail
- Interactive Data Analysis with Spark Shell
- Spark API for different File Formats & Compression Codecs
- Caching and Persistence
- Improving Spark Performance
- Spark Core Real-time Use cases
- Exploring Spark SQL
- SQL Realtime Usecases
- Exploring Spark Streaming
- Streaming Realtime Usecases
- Spark Machine Learning
- MLLib Realtime Usecases
- Kafka Messaging Queue
- Kafka Realtime Usecases
- Elastic Search Document Datastore
- Elastic Search Realtime Usecases
- Kibana
- Kibana Realtime Usecases
- End to End Project on realtime Fleet tracking analysis with spark streaming application, Kafka, NIFI, Cassandra, Elastic Search and Kibana Dashboard

# Use cases & Projects that we execute in this complete course

- Server Log Analysis
- Retail Banking
- Consumer - Product sales analysis
- Federated Data lake
- Movie Dataset Analysis
- Slowly Changing Dimension using Spark SQL
- Streaming Log files storage
- Consumer pricing usecase
- Twitter popular hashtag trending analysis
- Sensor data streaming pipeline
- Spam Filtering Analysis
- Uber Data Analysis
- End to End Project on realtime Fleet tracking analysis with spark streaming application, Kafka, NIFI, Cassandra, Elastic Search and Kibana Dashboard.

# BIG DATA, HADOOP & SPARK OVERVIEW

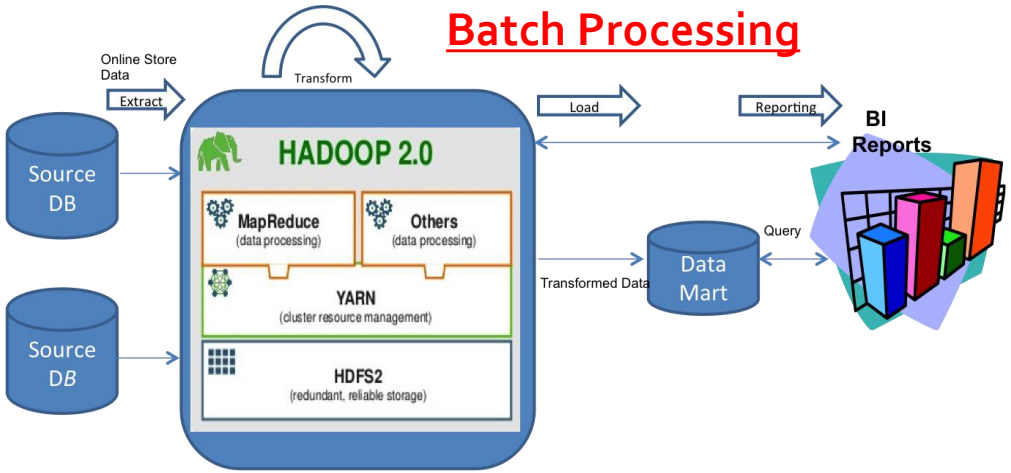
# BIG DATA

- Introduction
- Characteristics



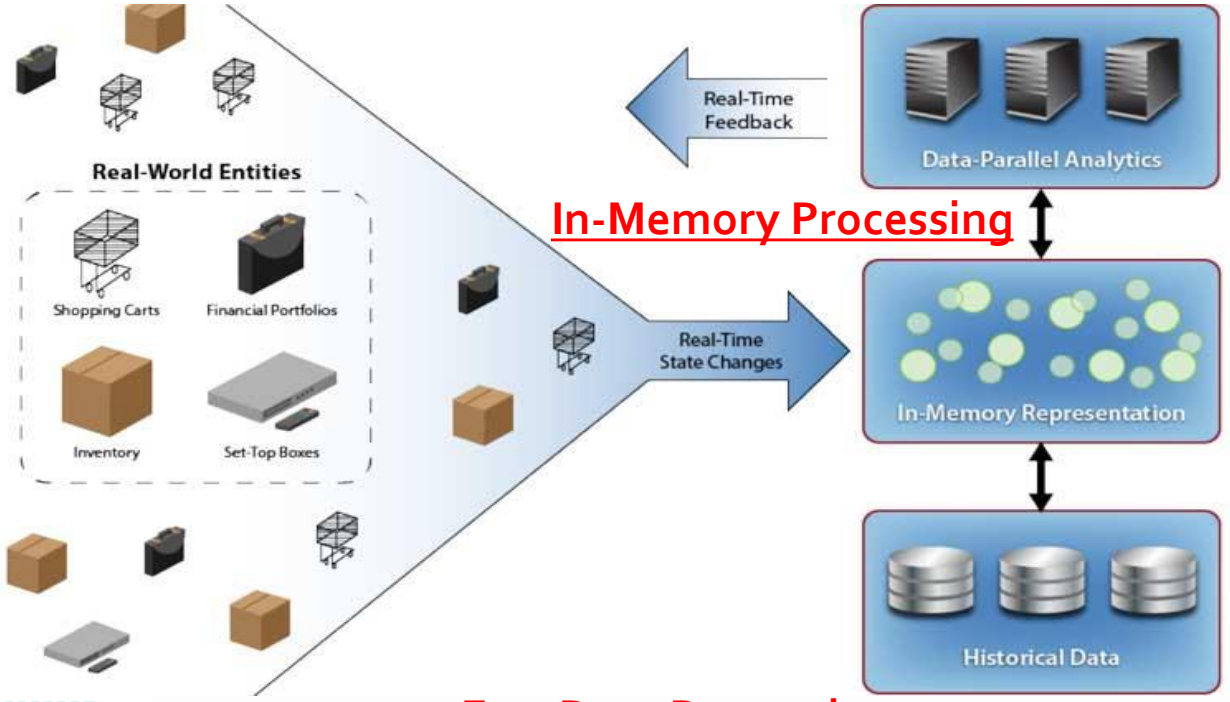
# Data Engineering Techniques

## Batch Processing

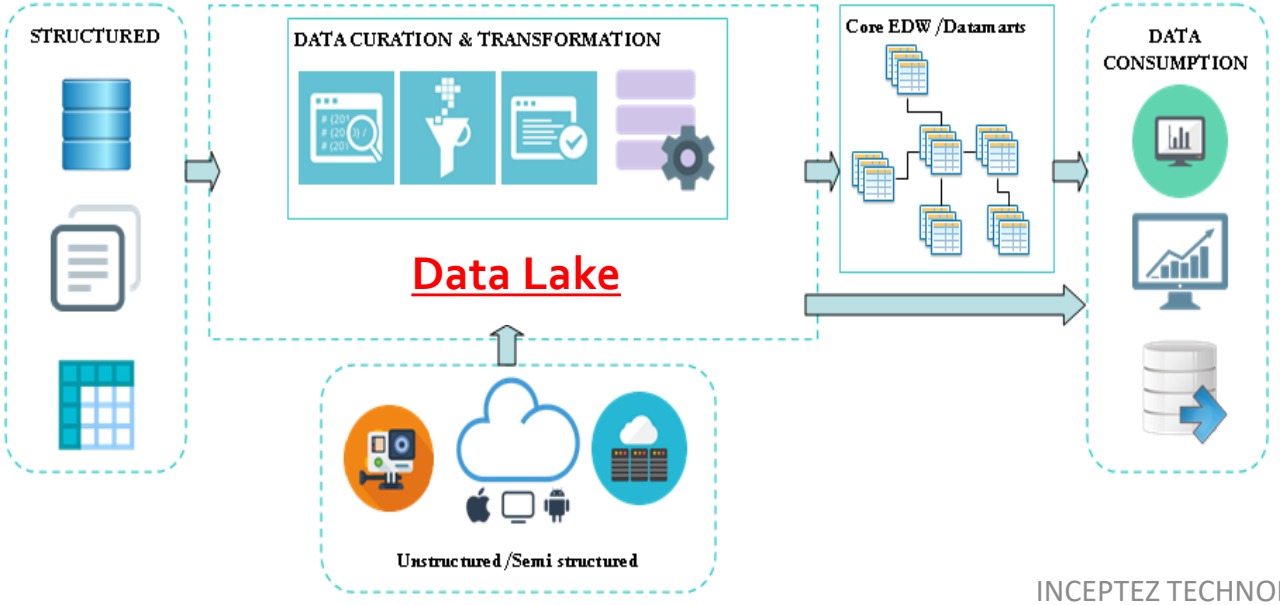



*The Big Data ETL*

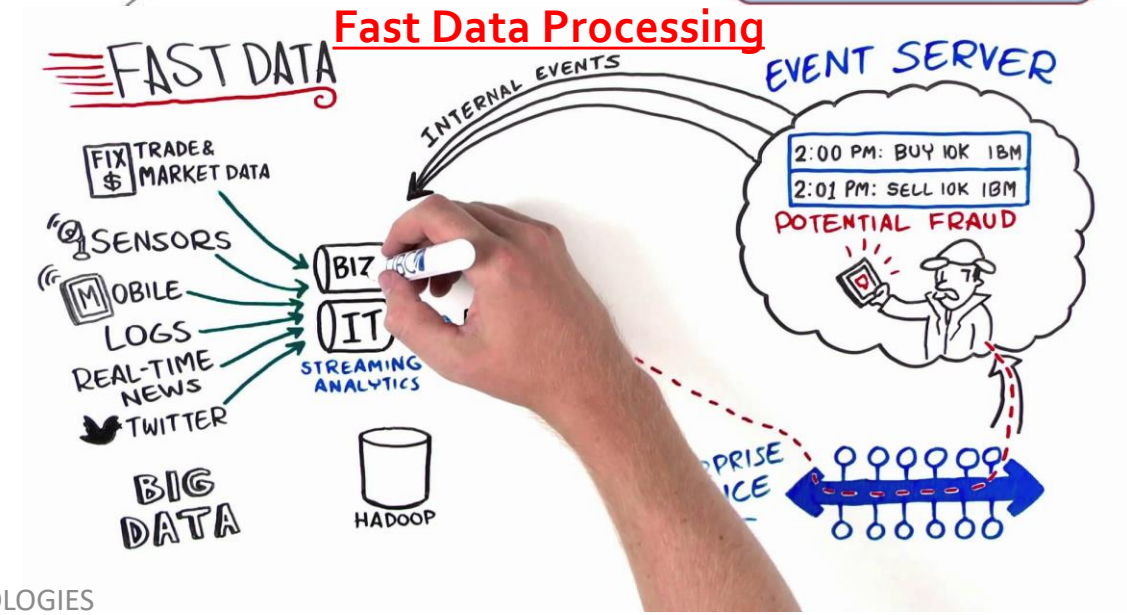
## In-Memory Processing




## Data Lake




## Fast Data Processing



# Distributed System

- Distributed storage
- Distributed processing
  - Resource Management
  - Distributed computation
- Network - connect to talk each other

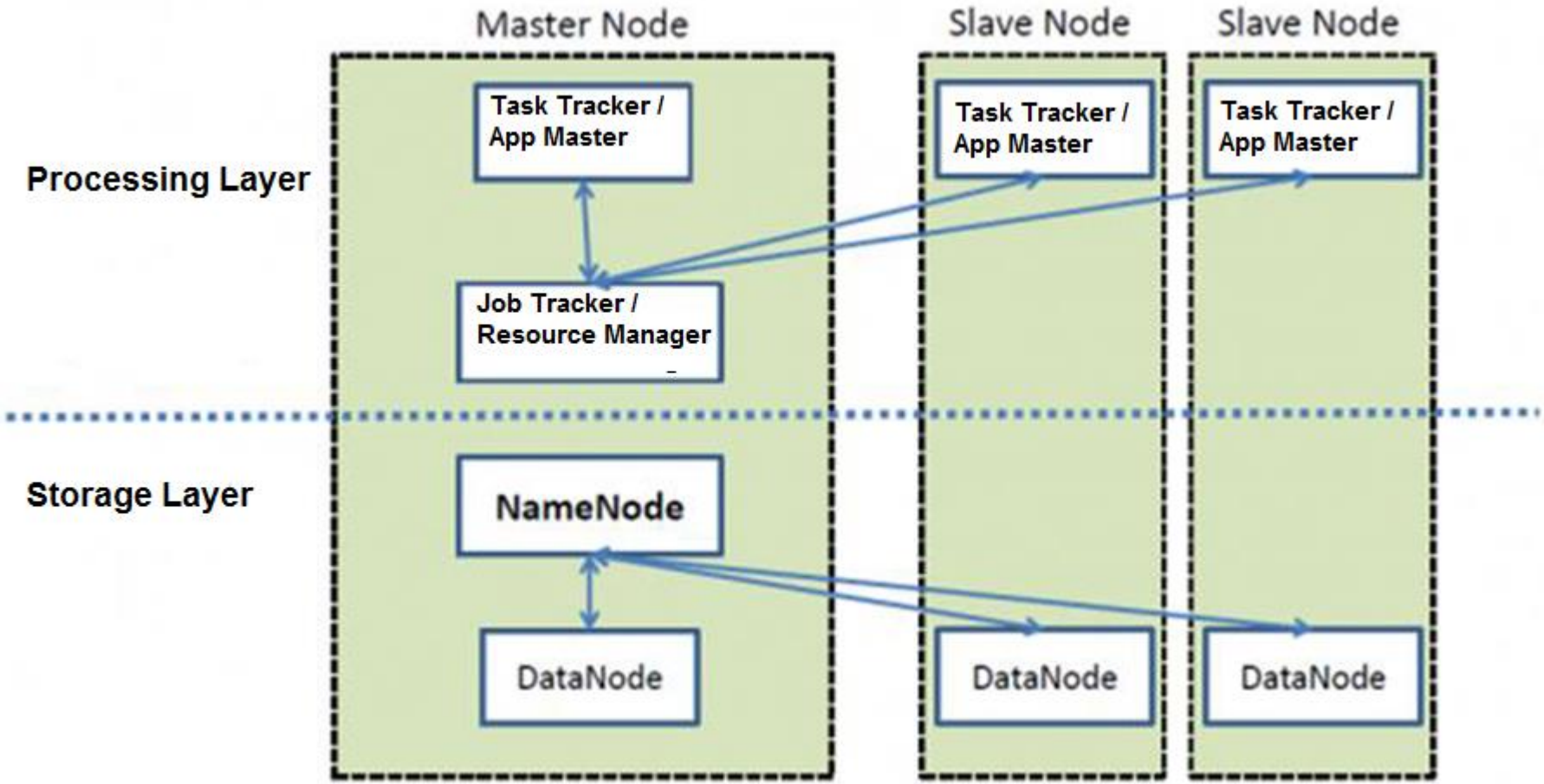
**Hadoop 1.0**

- HDFS - Distributed Storage
- Map/Reduce - Resource Management and distributed processing

**Hadoop 2.0**

- HDFS - Distributed storage
- Distributed Processing
  - YARN - Resource Management
  - Map/Reduce - Distributed Processing

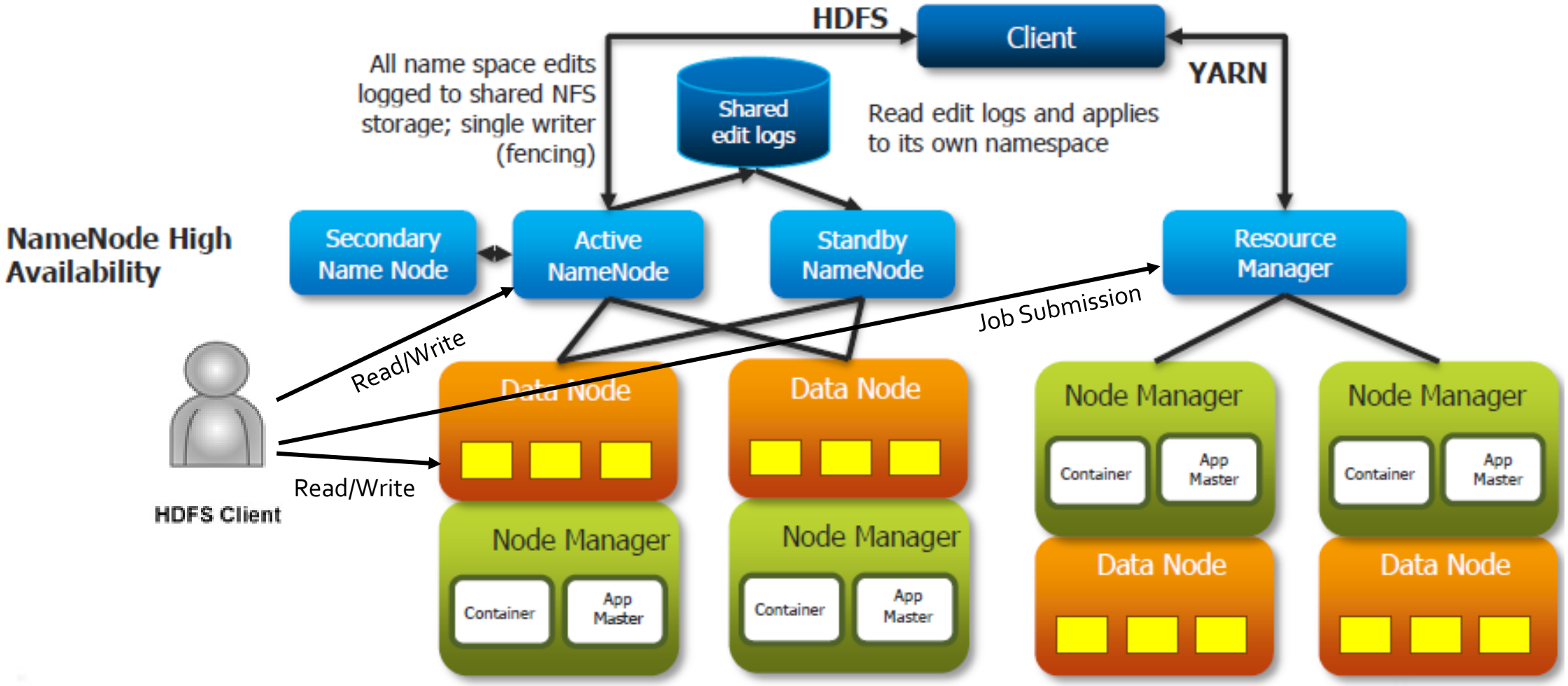
# High Level Architecture of Hadoop



# Hadoop Properties

- File System
- Immutability
- Replication
- Data Locality
- Fault Tolerance
- Scalability
- Batch Processing

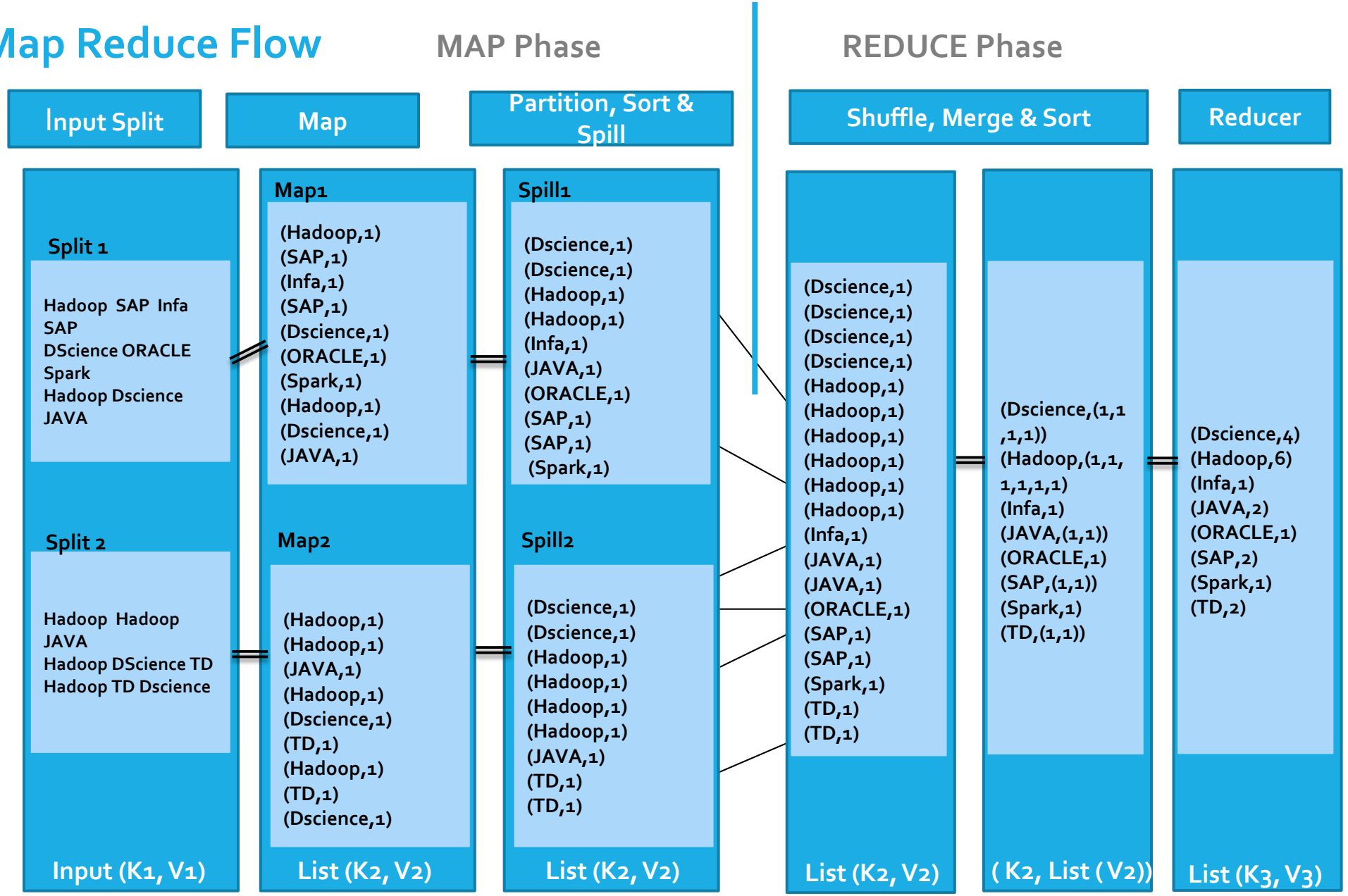
# Hadoop – A Quick Overview



# Map/Reduce Characteristics

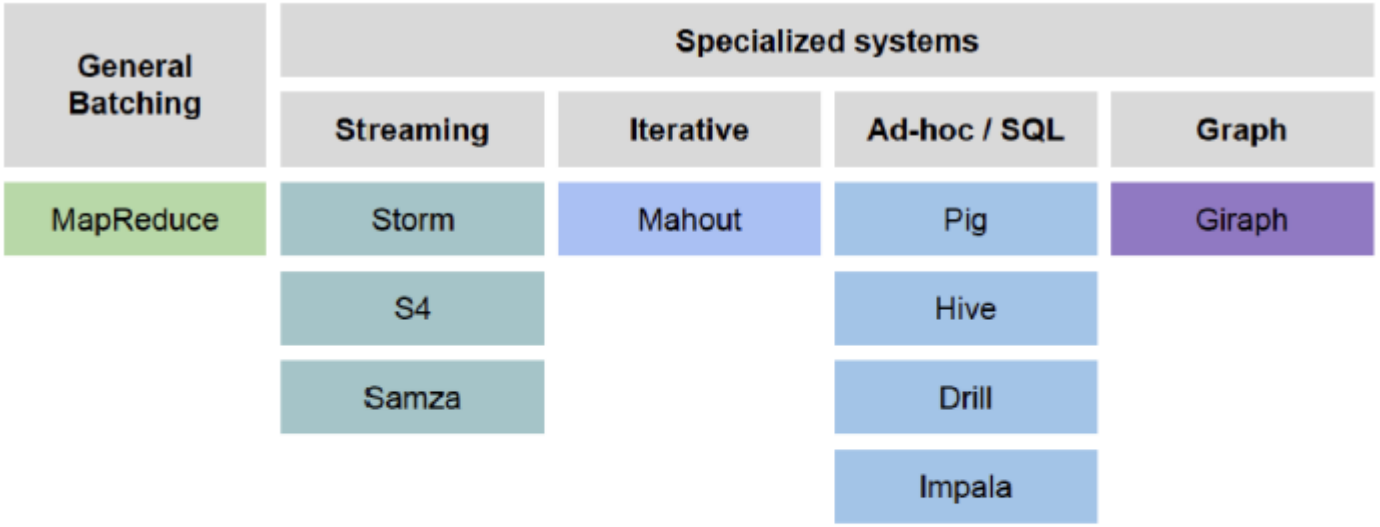
- Block Split /Input Split
- Driver
- Mapper
- Reducer
- Combiner
- Partitioner

# Map Reduce Flow



# Hadoop Ecosystems – No Unified Vision

- Sparse Modules
- More of Batch
- Diversity of Tools/APIs
- Huge coding efforts
- Heavily I/O Bounded
- High Latency



# Hadoop & Spark Approach

- **Hadoop introduced a radical new approach based on two key concepts**
  - Distribute the data when it is stored
  - Run computation where the data is
- **Spark takes this new approach to the next level**
  - Data is distributed in memory