Review : Hadoop vs. Typical DBMS

	Distributed DBs	Hadoop
Computing Model	- Notion of transactions - Transaction is the unit of work - ACID properties, Concurrency control	
Data Model	- Structured data with known schema - Read/Write mode	
Cost Model	- Expensive servers	
Fault Tolerance	- Failures are rare - Clever recovery mechanisms	
Key Characteristics	- Efficiency, Powerful, optimizations	

Review : Hadoop vs. Typical DBMS

	Distributed DBs	Hadoop
Computing Model	- Notion of transactions - Transaction is the unit of work - ACID properties, Concurrency control	- Notion of jobs - Job is the unit of work - No concurrency control
Data Model	- Structured data with known schema - Read/Write mode	- Any data format - ReadOnly mode
Cost Model	- Expensive servers	- Cheap commodity machines
Fault Tolerance	- Failures are rare - Clever recovery mechanisms	- Failures are common over thousands of machines - Simple fault tolerance
Key Characteristics	- Efficiency, powerful, optimizations	- Scalability, flexibility, fault tolerant

MapReduce Engine

- **JOB Tracker is the master node (runs with namenode)**
  - Receives "job" from user program
  - Decides on how many tasks will run (eg. number of mappers)
  - Decides on where to run each task (concept of locality)

- **TASK Tracker is the slave node (runs on each datanode)**
  - Receives the "task" from Job Tracker
  - Runs the task until completion (either map or reduce task)
  - Always in communication with Job Tracker reporting progress

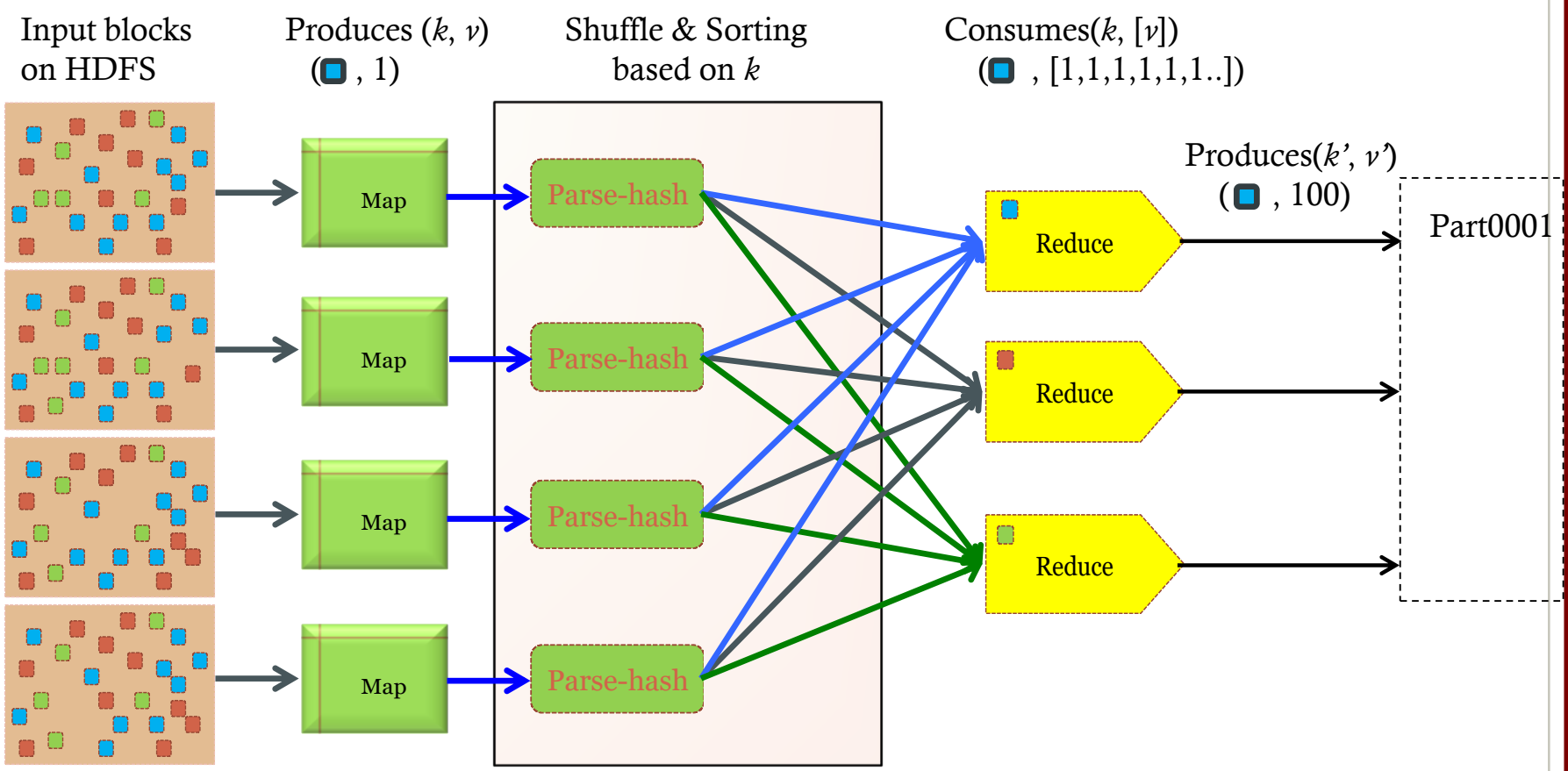
Key-Value Pairs

- Developer provides Mapper and Reducer functions
- Developer decides what is key and what is value

Key-Value Pairs

- **Mappers:** Run on a record-by-record
  - Consume <key, value> pairs
  - Produce <key, value> pairs
- **Reducers:** Run on a group-of-records with same key
  - Consume <key, <list of values>>
  - Produce <key, value>
- **Shuffling and Sorting:**
  - Hidden phase between mappers and reducers
  - Groups all similar keys from all mappers, sorts and passes them to a particular reducer in the form of <key, <list of values>>

Map-Reduce Execution Engine (Example 1: Color Count)



Users only provide the "Map" and "Reduce" functions