Business School
UNIVERSITY OF COLORADO DENVER

Information Systems Program

Module 5
Physical Design and Governance of Data Warehouses

Lesson 2: Scalable Parallel Processing Approaches

Lesson Objectives

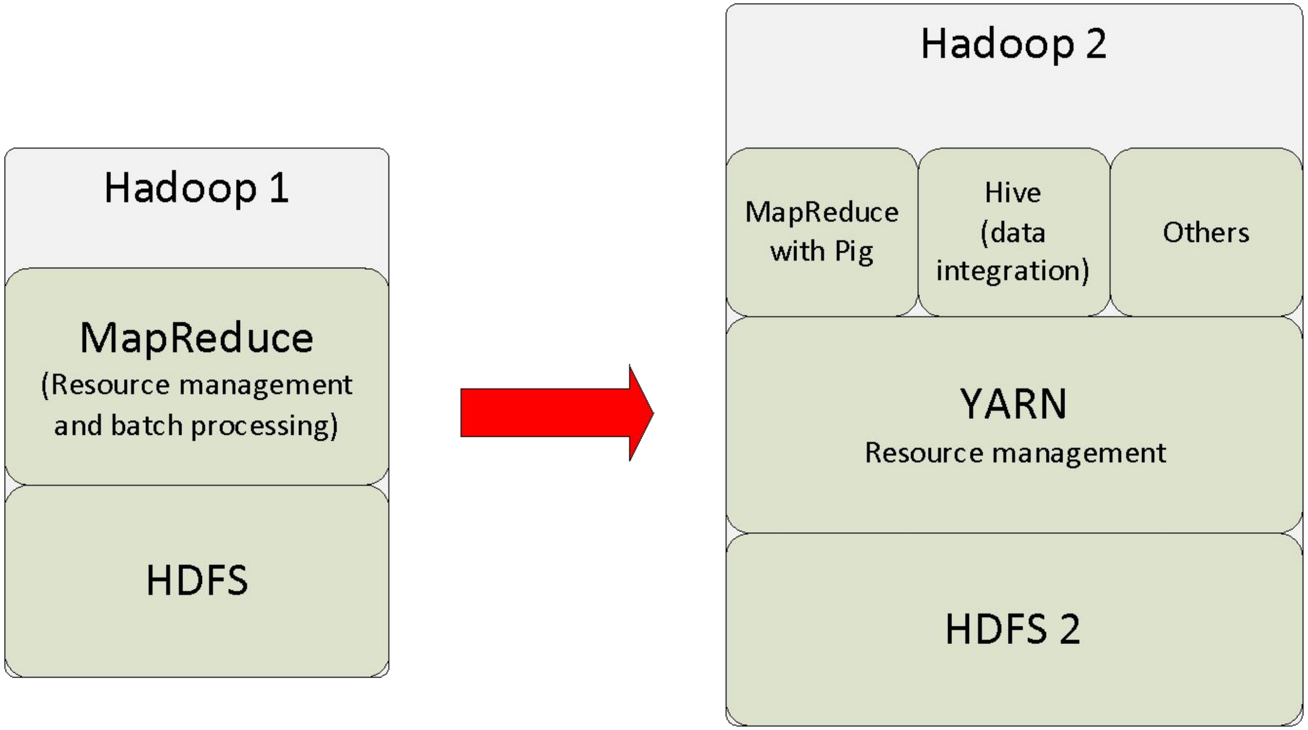
- Discuss importance of scalable parallel processing
- Explain Hadoop components
- Discuss usage of Hadoop for data integration

Timeline of Scalable Parallel Processing

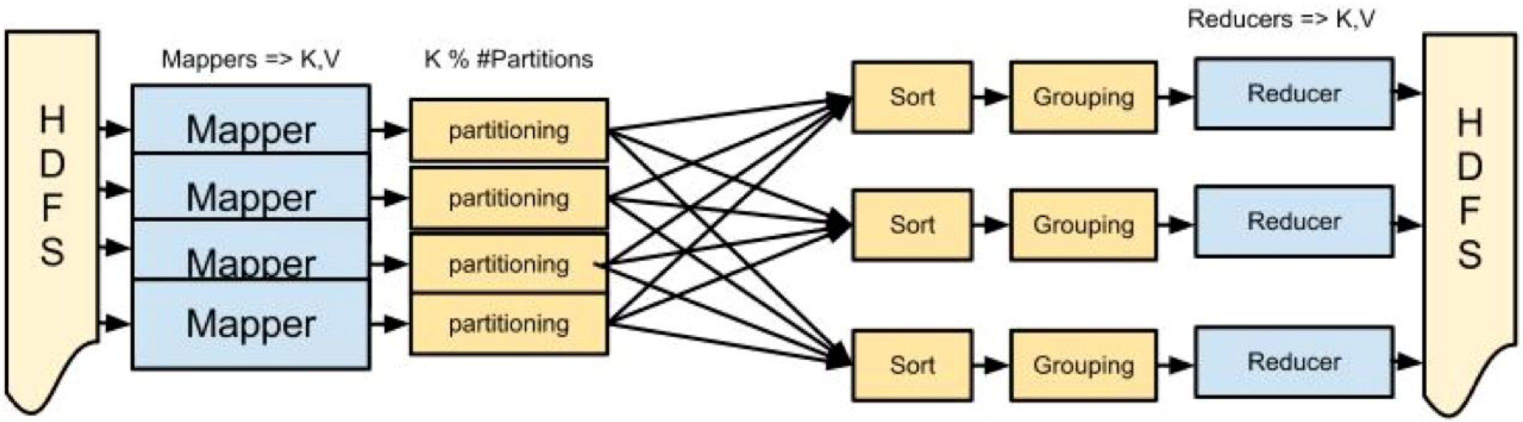
- Origins
  - Project Nutch (2002)
  - Google File System (2003)
  - Map Reduce paper (2004)
- Hadoop
  - Open source project (2005)
  - Cloudera founding (2009)
  - Hadoop 2 (2013)

- Open source project with commodity components
- API and services for parallel processing and job management
- Distributed file system
- Extensible for multiple task models

Hadoop Evolution



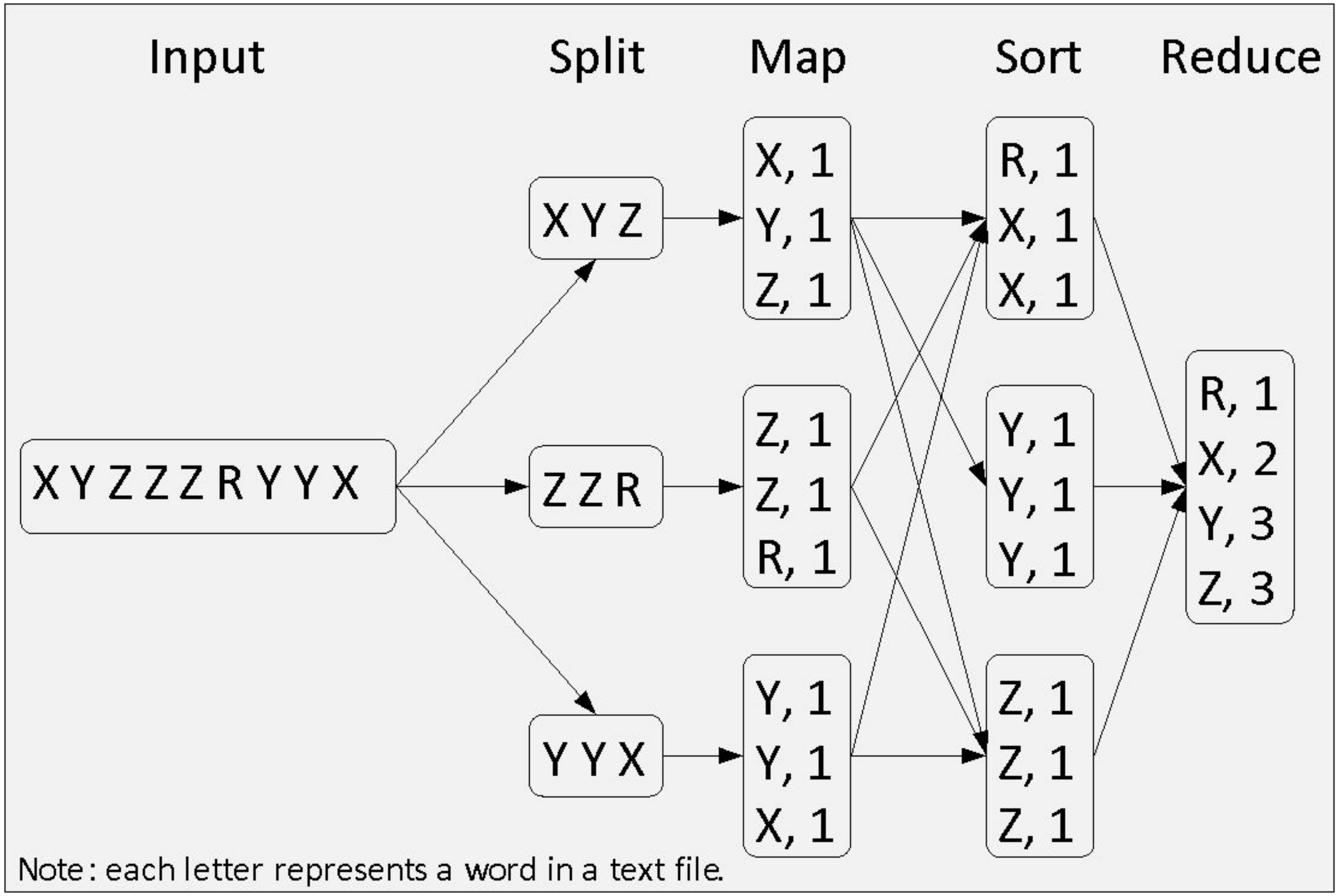
MapReduce Framework



The MapReduce Pipeline

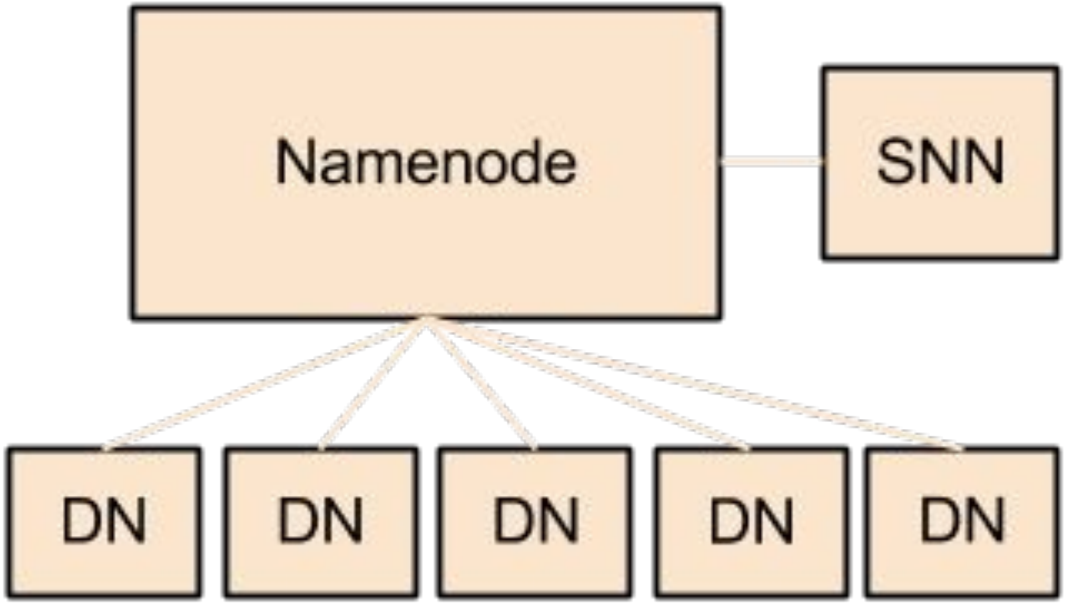
A mapper receives (Key, Value) & outputs (Key, Value)
A reducer receives (Key, Iterable[Value]) and outputs (Key, Value)
Partitioning / Sorting / Grouping provides the Iterable[Value] & Scaling

MapReduce Example



Note: each letter represents a word in a text file.

Distributed File System



Extensions to Big Data Processing

Improved performance and new tasks

Distributed, in-memory data sets in Apache Spark

Analytic query processing in Apache Hawq

Support for SQL queries, streaming analytics, data integration, and graph computations in Spark and Hawq

Summary

- Scalable, reliable parallel processing using commodity components
- Wide usage of Hadoop 2 open source project
- Growing importance of Hadoop for extended data integration