Lesson 2
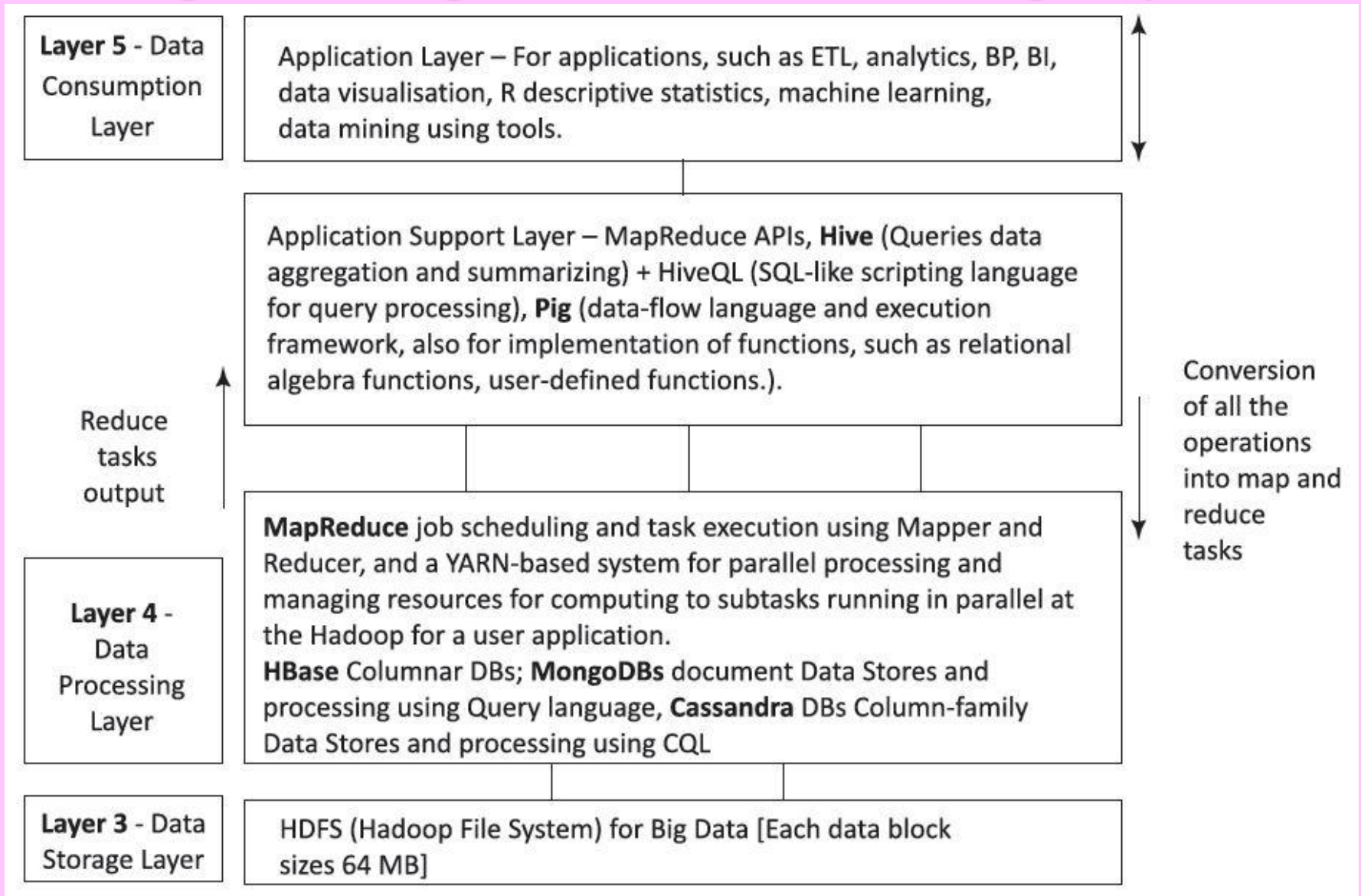# Introduction to Big Data Architecture Design Layers

Figure 4.1 Big Data architecture design layers

Layer 5 - Data Consumption Layer	Application Layer – For applications, such as ETL, analytics, BP, BI, data visualisation, R descriptive statistics, machine learning, data mining using tools.

Application Support Layer – MapReduce APIs, **Hive** (Queries data aggregation and summarizing) + HiveQL (SQL-like scripting language for query processing), **Pig** (data-flow language and execution framework, also for implementation of functions, such as relational algebra functions, user-defined functions.).

Reduce tasks output

Conversion of all the operations into map and reduce tasks

Layer 4 - Data Processing Layer

MapReduce job scheduling and task execution using Mapper and Reducer, and a YARN-based system for parallel processing and managing resources for computing to subtasks running in parallel at the Hadoop for a user application.
HBase Columnar DBs; **MongoDBs** document Data Stores and processing using Query language, **Cassandra** DBs Column-family Data Stores and processing using CQL

Layer 3 - Data Storage Layer

HDFS (Hadoop File System) for Big Data [Each data block sizes 64 MB]

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education (India)

Big Data Architecture Design Layers

<u>Refer Figure 4.1</u>

- Application and Application Support APIs

- Codes Converted and inputs to Mapper-Reducer Tasks

- Data Processing using Mapper- Reducer Tasks

- Data store units, (data, in-processing results, and Output results)

Big Data Architecture Design Layers

Refer Figure 4.1 Data Flow

- Data Store units

- Data Processing, Mapper and Reducer Tasks

- Output of Reducer Tasks to─

- Application Support APIs

- Application

Layers 1 and 2

Refer Figure 1.2

- Layer 1 ─ Identification of internal and external Sources of Data

- Layer 2 ─ Data ingestion and acquisition

Layer 3: Data Store Layer (Distributed Data)

- Figure 4.1:

- Hadoop Distributed File System (HDFS) where clusters store the data, in-processing results for the application tasks

Data Store

- Figure 4.1

- HDFS

- HBase Columnar DBs;

- MongoDBs Document Data Stores and processing using Query language Cassandra DBs Column-family Data Stores and processing using CQL

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education (India)

Layer 4: Data Processing Layer

- Figure 4.1

- Resource Management and

- Processing Framework

Resources Management

- Making available resources of CPU, RAM and network in the Hadoop clusters for the multiple application subtasks and tasks (Job Tracker Daemon/YARN)

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education   (India)

Yarn

- A component of Hadoop, providing:

- Resources management using multiple machines (data nodes)

- Running and scheduling of the parallel programs for map and reduce tasks

- Allocating parallel processing resources for computing subtasks

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education   (India)

Processing Framework

- Figure 4.1

- Mapping, aggregation with shuffle, sort or merge in environment using map and reduce tasks

- Reducer (JobTracker/Task Tracker Daemons) at slave nodes in the clusters

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education (India)

Layer 5: Application Support Layer APIs

- Figure 4.1

- Application tasks/subtasks coded in MapReduce APIs, HBase or Hive/Pig projects

- Hive/Pig easier programming models

- Codes in Hive/Pig compiles into MapReduce tasks

Layer 6: Data Consumption Application Support Layer APIs

- Figure 4.1

- Application Layer– for applications, ETL, Analytics, BP, BI, Data Visualization, R- Descriptive Statistics, Machine learning, Data mining using tools

Hive

- **Hive** (Queries data aggregation and summarizing) +

- HiveQL(SQL-like scripting language for the Query Processing)

"Big Data Analytics ", Ch.04 L02: MapReduce, Hive and Pig
Raj Kamal, and Preeti Saxena © McGraw-Hill Education   (India)

Pig

- A data-flow language and execution framework, also for implementation of functions, such as relational algebra functions, user-defined functions.

Summary

We learnt :

- Application and Application Support Layers APIs

- Programming in MapReduce or HBase/Hive/Pig

Summary

- Data Processing Framework

- Resource Management

- Data Store: HDFS, MongoDBs, Cassandra DBs

End of Lesson 2 on
Introduction to Big Data
Architecture Design Layers