Lesson 7

Applications and Big Data analytics using Spark

Big Data Architecture Design Layers

Refer Figure 4.1

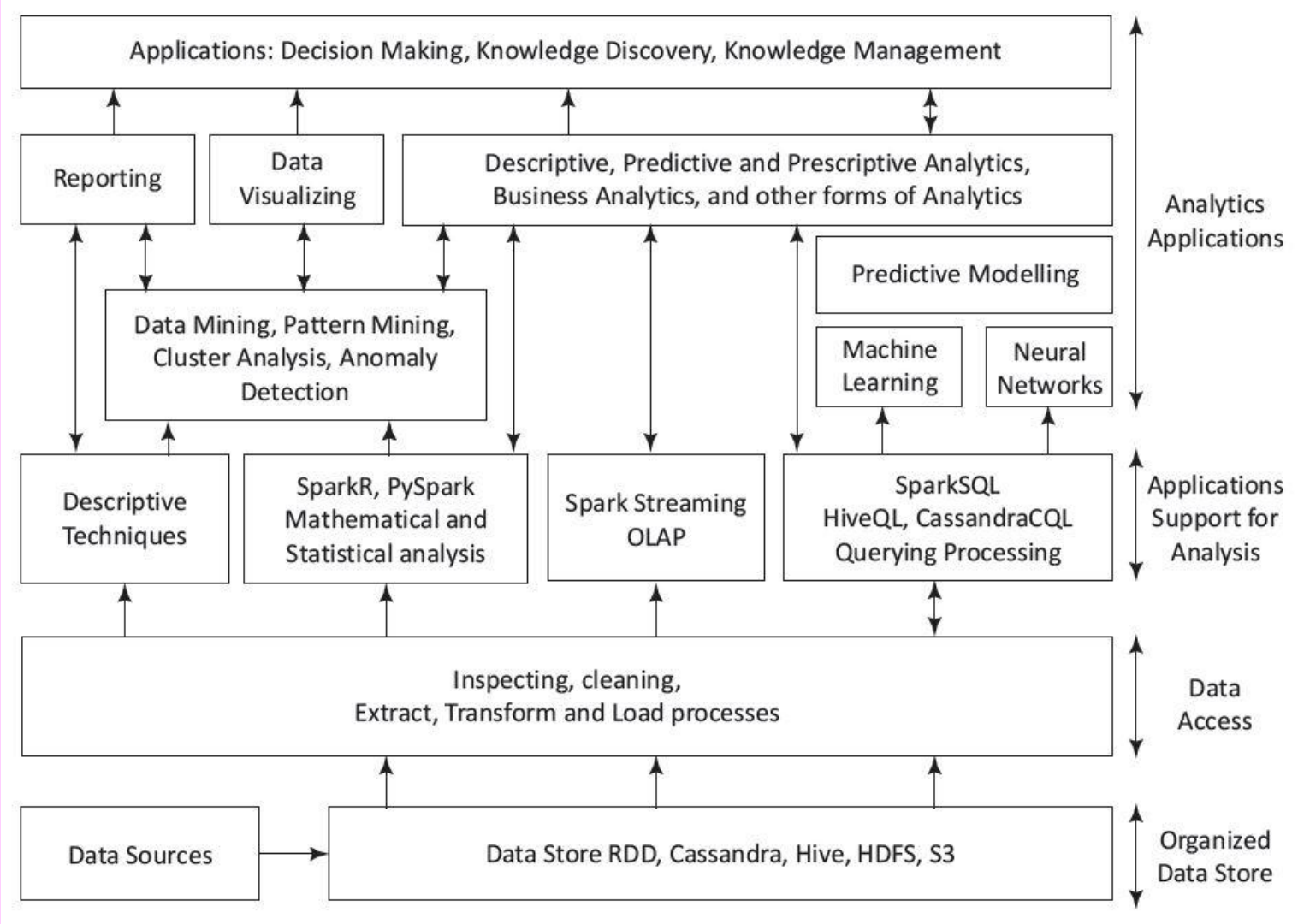
- Application and Application Support APIs
- Codes Converted and inputs to Mapper-Reducer Tasks
- Data Processing using Mapper- Reducer Tasks
- Data store units, (data, in-processing results, and Output results)

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

Figure 5.10 Processing framework for applications and Big Data analytics using Spark



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

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)

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

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

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)

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 7 on

Applications and Big Data analytics using Spark