Lesson 5

Data storage and analysis, Comparison between traditional systems, and Approaches for Big Data storage and analytics

Traditional systems

- Structured Data
- RDBMS data, such as MySQL DB2, enterprise server and data warehouse
- SQL─ a language for managing the RDBMS Relational database examples are MySQL PostGreSQL Oracle database, Informix, IBM DB2 and Microsoft SQL server.

Traditional Systems

- Semi-Structured Data

- XML and JSON

- A comma-separated values (CSV) file

Enterprise Data

- Enterprise Data-Store Server

- Data Warehouse

- Enterprise data warehouse store the databases, and data stores after integration, using tools from number of sources

Big Data Systems

- NoSQL databases considered as semi-structured data

- Big Data Store uses NoSQL. NOSQL stands for No SQL or Not Only SQL

- The stores do not integrate with applications using SQL

Figure 1.6 Steps 1 to 5 in Enterprise data integration and management with Big-Data for high performance computing
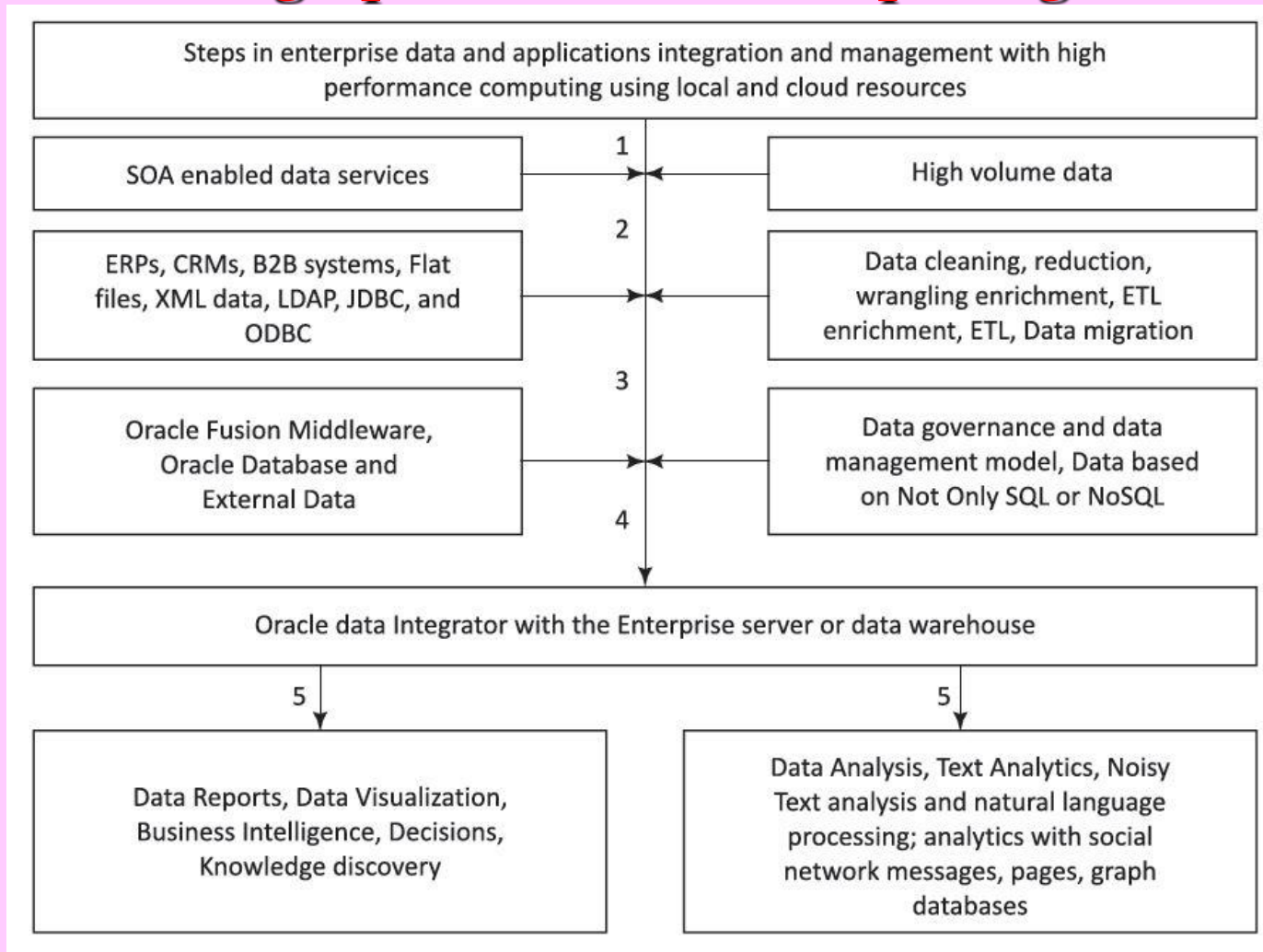
Steps in enterprise data and applications integration and management with high performance computing using local and cloud resources

	1	
SOA enabled data services		High volume data

	2	
ERPs, CRMs, B2B systems, Flat files, XML data, LDAP, JDBC, and ODBC		Data cleaning, reduction, wrangling enrichment, ETL enrichment, ETL, Data migration

	3	
Oracle Fusion Middleware, Oracle Database and External Data		Data governance and data management model, Data based on Not Only SQL or NoSQL

4

Oracle data Integrator with the Enterprise server or data warehouse

5   5

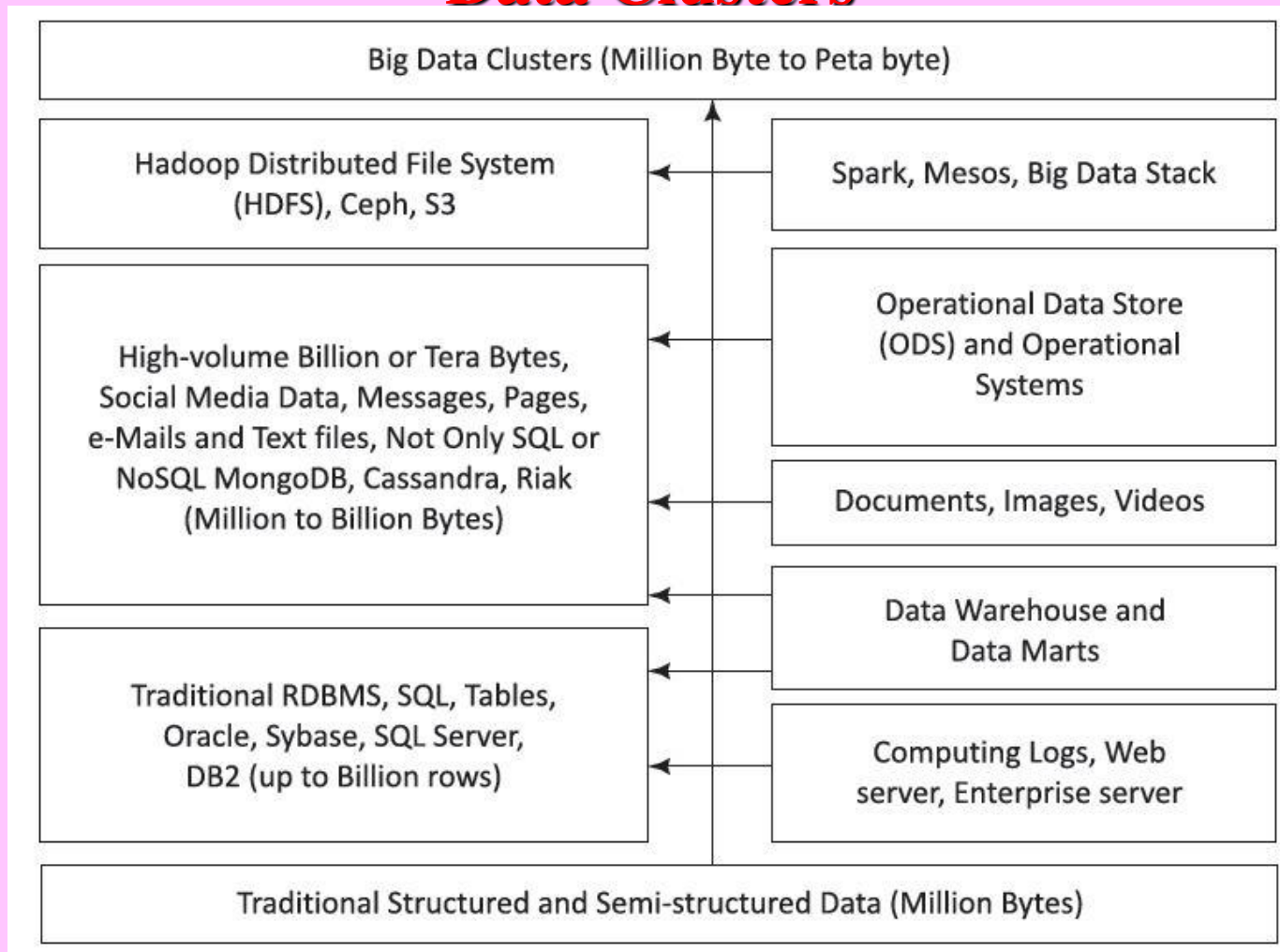| Data Reports, Data Visualization, Business Intelligence, Decisions, Knowledge discovery	Data Analysis, Text Analytics, Noisy Text analysis and natural language processing; analytics with social network messages, pages, graph databases

Figure 1.7 Coexistence of traditional server data, NoSQL and Hadoop, Spark and compatible Big Data Clusters

Big Data Clusters (Million Byte to Peta byte)	
Hadoop Distributed File System (HDFS), Ceph, S3	Spark, Mesos, Big Data Stack
High-volume Billion or Tera Bytes, Social Media Data, Messages, Pages, e-Mails and Text files, Not Only SQL or NoSQL MongoDB, Cassandra, Riak (Million to Billion Bytes)	Operational Data Store (ODS) and Operational Systems
	Documents, Images, Videos
	Data Warehouse and Data Marts
Traditional RDBMS, SQL, Tables, Oracle, Sybase, SQL Server, DB2 (up to Billion rows)	Computing Logs, Web server, Enterprise server
Traditional Structured and Semi-structured Data (Million Bytes)	

"Big Data Analytics ", Ch.01 L05: Introduction To ... Big Data Analytics Raj Kamal and Preeti Saxena, © McGraw-Hill Higher Edu. India

Big Data Platform Data management, storage and Analytics

1. New innovative non-traditional methods of storage, processing and analytics

2. Distributed Data Stores

3. Creating scalable as well as elastic virtualized platform (cloud computing)

Big Data Platform Data management, storage and Analytics

4. Huge volume of Data Stores

5. Massive parallelism

6. High speed networks

7. High performance processing, optimization and tuning

8. Data management model based on Not Only SQL or NoSQL

"Big Data Analytics ", Ch.01 L05: Introduction To ... Big Data Analytics
Raj Kamal and Preeti Saxena, © McGraw-Hill Higher Edu. India

Big Data Platform Data management, storage and Analytics

9. In-memory data column-formats transactions processing or dual in-memory data columns as well as row formats for OLAP and OLTP

10. Data retrieval, mining, reporting, visualization and analytics

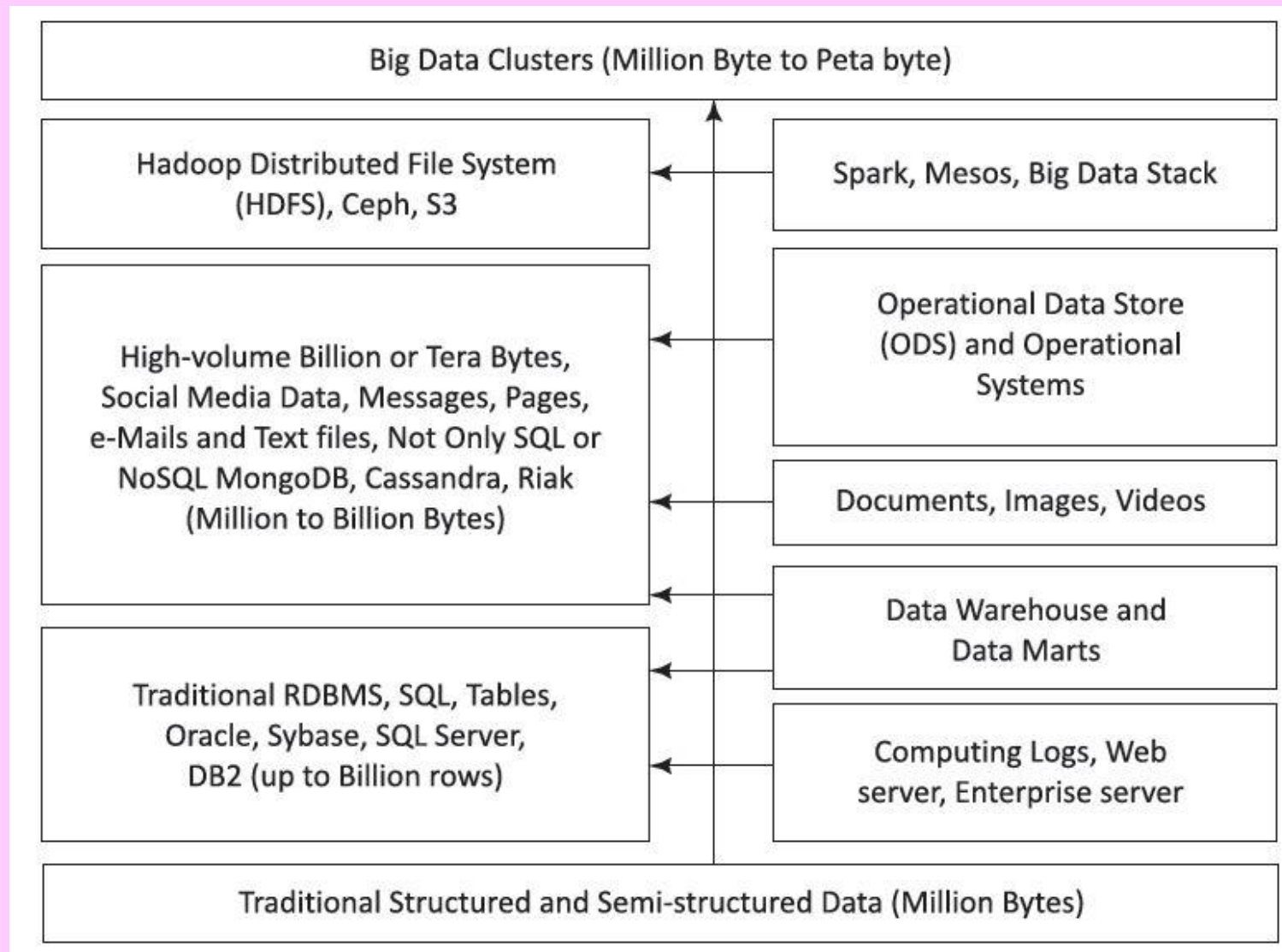# Big Data Platform Data management, storage and Analytics

11. Graph databases to enable analytics with social network messages, pages and data analytics

12. Machine learning or other approaches

Big Data Platform Data management, storage and Analytics

13. Big data sources: Data storages, data warehouse, Oracle Big Data, MongoDB NoSQL, Cassandra NoSQL

14. Data sources: Sensors, Audit trail of Financial transactions data, external data such as Web, Social Media, weather data, health records data.

Figure 1.8 Hadoop based Big Data environment

Big Data Clusters (Million Byte to Peta byte)

Hadoop Distributed File System (HDFS), Ceph, S3	Spark, Mesos, Big Data Stack
High-volume Billion or Tera Bytes, Social Media Data, Messages, Pages, e-Mails and Text files, Not Only SQL or NoSQL MongoDB, Cassandra, Riak (Million to Billion Bytes)	Operational Data Store (ODS) and Operational Systems
	Documents, Images, Videos
	Data Warehouse and Data Marts
Traditional RDBMS, SQL, Tables, Oracle, Sybase, SQL Server, DB2 (up to Billion rows)	Computing Logs, Web server, Enterprise server

Traditional Structured and Semi-structured Data (Million Bytes)

"Big Data Analytics ", Ch.01 L05: Introduction To ... Big Data Analytics
Raj Kamal and Preeti Saxena, © McGraw-Hill Higher Edu. India
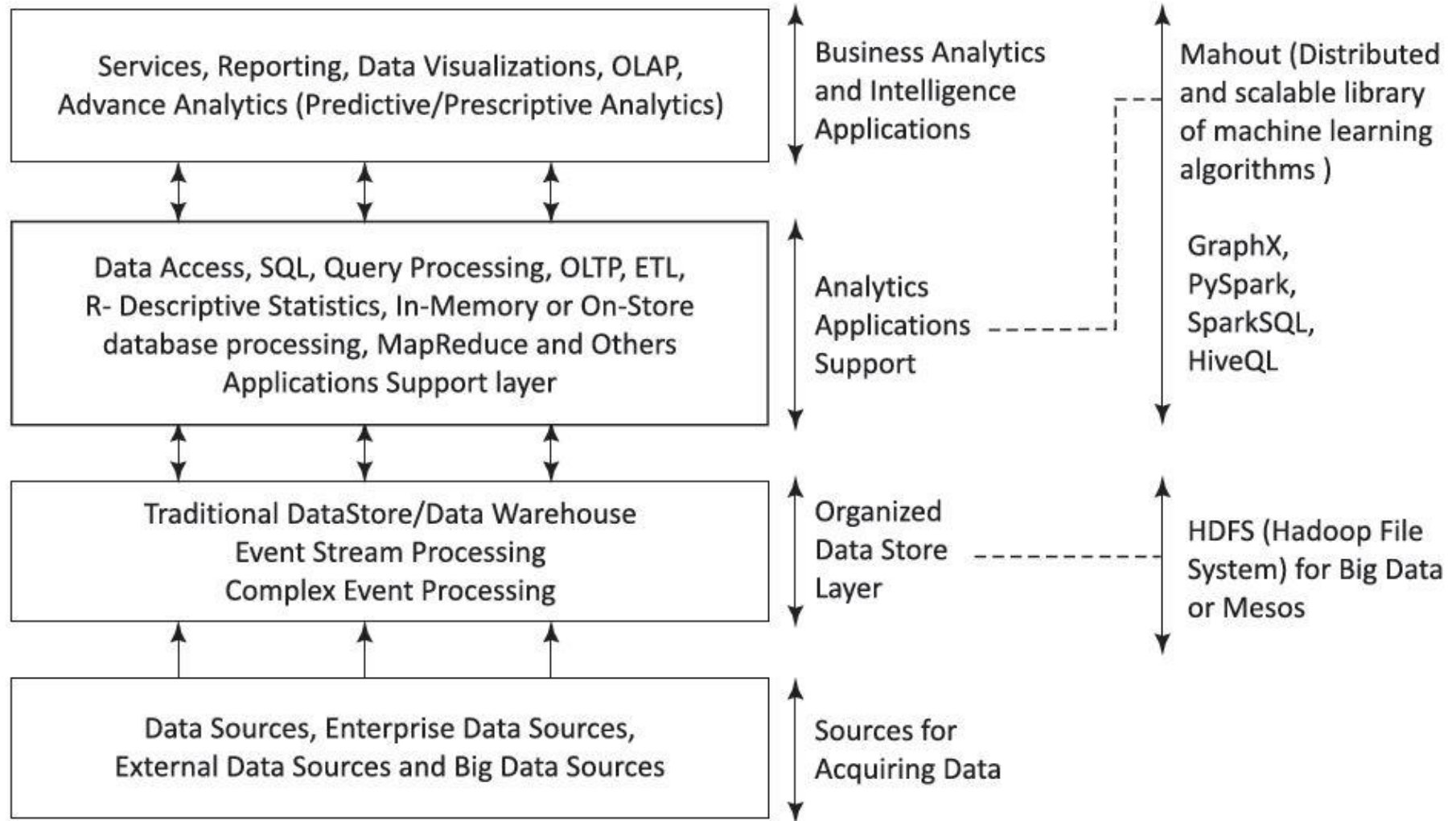
Big Data Stack

- A stack consists of a set of software components and data store units. Applications, machine-learning algorithms, analytics and visualization tools

Big Data Stack

- Use Big Data Stack (BDS) at a cloud service, such as Amazon EC2, Azure or private cloud. The stack uses cluster of high performance machines

Figure 1.9 Traditional and Big Data analytics architecture reference model

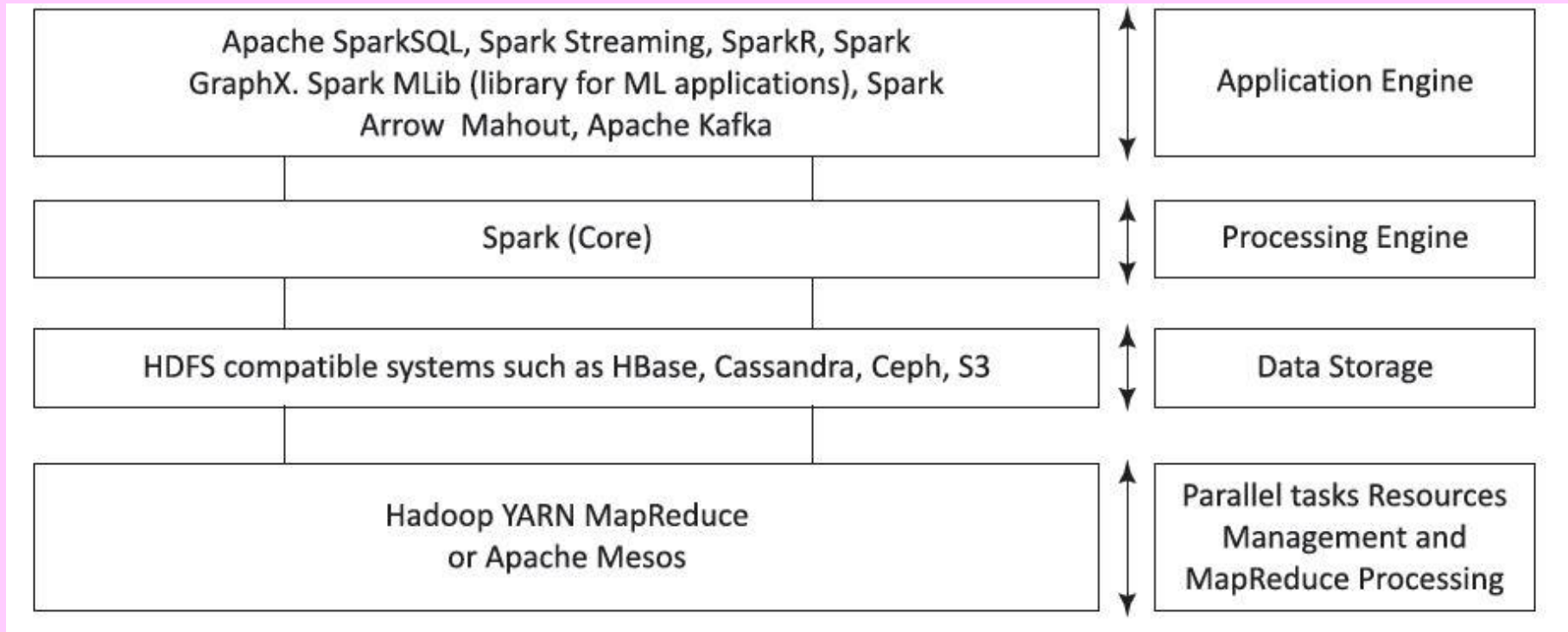|		
Services, Reporting, Data Visualizations, OLAP, Advance Analytics (Predictive/Prescriptive Analytics)	Business Analytics and Intelligence Applications	Mahout (Distributed and scalable library of machine learning algorithms)
Data Access, SQL, Query Processing, OLTP, ETL, R- Descriptive Statistics, In-Memory or On-Store database processing, MapReduce and Others Applications Support layer	Analytics Applications Support	GraphX, PySpark, SparkSQL, HiveQL
Traditional DataStore/Data Warehouse Event Stream Processing Complex Event Processing	Organized Data Store Layer	HDFS (Hadoop File System) for Big Data or Mesos
Data Sources, Enterprise Data Sources, External Data Sources and Big Data Sources	Sources for Acquiring Data	

Berkeley Data Analytics Stack (BDAS)

- Consists of data processing, data management and resource management layers

- 1. Data processing software component provides in-memory processing which processes the data efficiently across the frameworks.

BDAS

2. Data processing combines batch, streaming and interactive computations.

3. Resource management software component provides for sharing the infrastructure across various frameworks

"Big Data Analytics ", Ch.01 L05: Introduction To ... Big Data Analytics
Raj Kamal and Preeti Saxena, © McGraw-Hill Higher Edu. India

Figure 1.10 Four layers architecture for Big Data Stack

Apache SparkSQL, Spark Streaming, SparkR, Spark GraphX. Spark MLib (library for ML applications), Spark Arrow Mahout, Apache Kafka	Application Engine
Spark (Core)	Processing Engine
HDFS compatible systems such as HBase, Cassandra, Ceph, S3	Data Storage
Hadoop YARN MapReduce or Apache Mesos	Parallel tasks Resources Management and MapReduce Processing

Summary

We learnt :

- Traditional Data Systems: RDBMS data, such as MySQL DB2, enterprise server and data warehouse

- Big Data Systems

- Traditional and Big Data analytics architecture reference model

Summary

We learnt :

- Hadoop based Big Data environment Big Data Systems

- Berkeley Data Analytics Stack (BDAS)

End of Lesson 5 on
Data storage and analysis, Comparison between traditional systems, and Approaches for Big Data storage and analytics

"Big Data Analytics ", Ch.01 L05: Introduction To ... Big Data Analytics
Raj Kamal and Preeti Saxena, © McGraw-Hill Higher Edu. India