Lesson 11
Big Data Analytics

Recall RDBMS

- A relational database is a collection of data into multiple tables which relates to each other through special fields, called keys (primary key, foreign key and unique key)
- RDBMS: a management system for the relational databases, (Creation, connect, insertion, append, replace, …)

Recall SQL

- **A** Structured Query Language
- A language for viewing or changing (update, insert or append or delete) databases
- A language for data querying the databases

Recall SQL

- A language for data access control,
- A language for schema creation and modifications
- Also a language for managing the RDBMS
- Also a language that can embed into other languages

Recall NoSQL (also called Not-Only SQL)

- A Class of non-relational data storage systems, flexible data models multiple schemas
- Class consisting of un-interpreted key and value or ‘the big hash table’
- Class consisting of unordered keys and using the JSON, for example in MongoDb

No SQL

- Class consisting of ordered keys and semi-structured data storage systems
- HBase
- BigTable
- Cassandra (used in Facebook and Apache)

Big Data: Data of high volume

- Big data is data of high volume, variety and velocity, and may also include veracity
- Volume means data received from number of sources of data
- Includes data sets with sizes beyond the ability of commonly used software tools to acquire, manage and process data within a tolerable elapsed time

Big Data: Data of high Variety

- Variety means structured as well as unstructured data in different formats
- variety of data on which no SQL (Structured Query Language) applicable
- The multi-structured data compared to RDMS which maintains more structured data

Big Data: of higher velocity

- Velocity means data received with higher rates due to use of number of sources of data
-

Big Data: Data of high veracity

- May include Veracity
- Variation in data quality for analytics

Bigtable

- NoSQL Big Data database
- Maps two arbitrary string values into an associated arbitrary byte array. One is used as row key and other as column key
- Time stamp also associates in three-dimensional mapping

Bigtable

- Mapping of the row and column keys unlike a relational database
- Mapping can be considered as a sparse (thinly dispersed or scattered) and for a distributed multi-dimensional sorted map
- The table can scale to 100s to 1000s of distributed computing nodes
- Ease of adding more nodes

Big Data for Analytics

- Data, filtered data after removing—anomalous data, non-standard and not cross referencing data
- The analytics need the trustable

Big Data Analytics Tools

- The open source software Hadoop and MapReduce from Apache Software
- Enable the storage
- Enable analysis of the massive amounts of data

Hadoop

- An open-source framework for accesses to data in sequential manner
- Performs batch processing
- A new data set results from input data set that also processes sequentially
- Hadoop file system (HDFS)

HBase

- Database for big data
- Data access– random access
- Provides fast look-up from large tables
- Small access latency
- Database using big hash tables
- Considered similar to Google's BigTable.

HiveQL for Big Data analytics

- HiveQL, a SQL like scripting language software
- Used in Hadoop ecosystem (a collection of related entities and certain processes that link to Hadoop components)
- (Hive is word derived from structure for housing domesticated honeybees)

MapReduce

- MapReduce is programming model
- A core of Hadoop
- Large data sets process onto a cluster of nodes using MapReduce
- Same node runs the algorithm using the data sets at HDFS and processing is at that node itself

Mahout

- Distributed and Scalable Library of machine learning algorithms

Berkeley Data Analytics Stack

- Stack three layers
- Data processing, data management, and resource management layers
- HDFS
- HiveQL
- Mahout

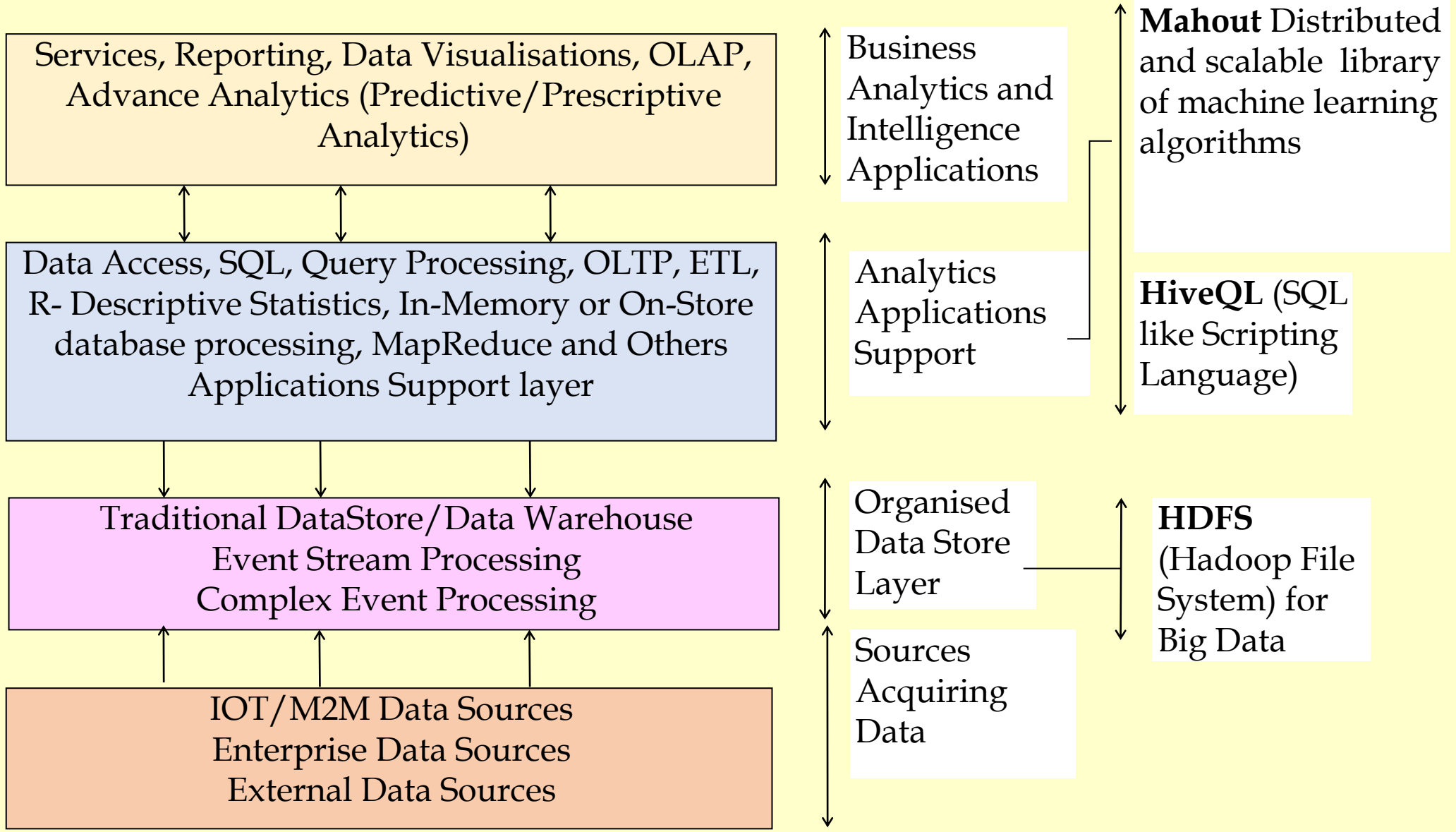

Fig. 5.6 Berkeley Data Analytics Stack Architecture

Summary

We learnt

- Big Data
- Volume, Variety, Velocity and Veracity
- Bigtable
- Analytics tools for Big Data
- Hadoop, HBase, MapReduce
- Mahout, HiveQL and HDFS

End of Lesson 11 on Big Data Analytics