BigData Analytics

Unit 5:
Exploring NoSQL Query Language

Exploring NoSQL Query Language

5.1. Introduction and its features

5.2. Data Types and Operators

5.3. CRUD Operations

5.4. Built-in Functions and Aggregate Functions

5.5. Import and Export Files

CE: 5.1. Introduction and its features

CE: 5.1.1. What is MongoDB?

- MongoDB is…
  - ✓ Cross-platform.
  - ✓ Open source.
  - ✓ Non-relational.
  - ✓ Distributed.
  - ✓ NoSQL.
  - ✓ Document-oriented data store.

- MongoDB is a free and open-source cross-platform document-oriented database.

- It is Classified as a NoSQL database.

- It avoids the traditional table-based relational database structure in favor of JSON - like documents with dynamic schemas.

CE: 5.1.2. Why MongoDB?

- Because…
  - ✓ Document oriented
  - ✓ High performance
  - ✓ Fast in-place updates
  - ✓ Replication
  - ✓ High availability (is fault tolerant)
  - ✓ Easy scalability
  - ✓ Rich query language
  - ✓ Full index support (faster queries)
  - ✓ Auto sharding
    (Distributes data across a cluster)

CE: 5.1.3. Query Language

Which Query Language is used by MongoDB?

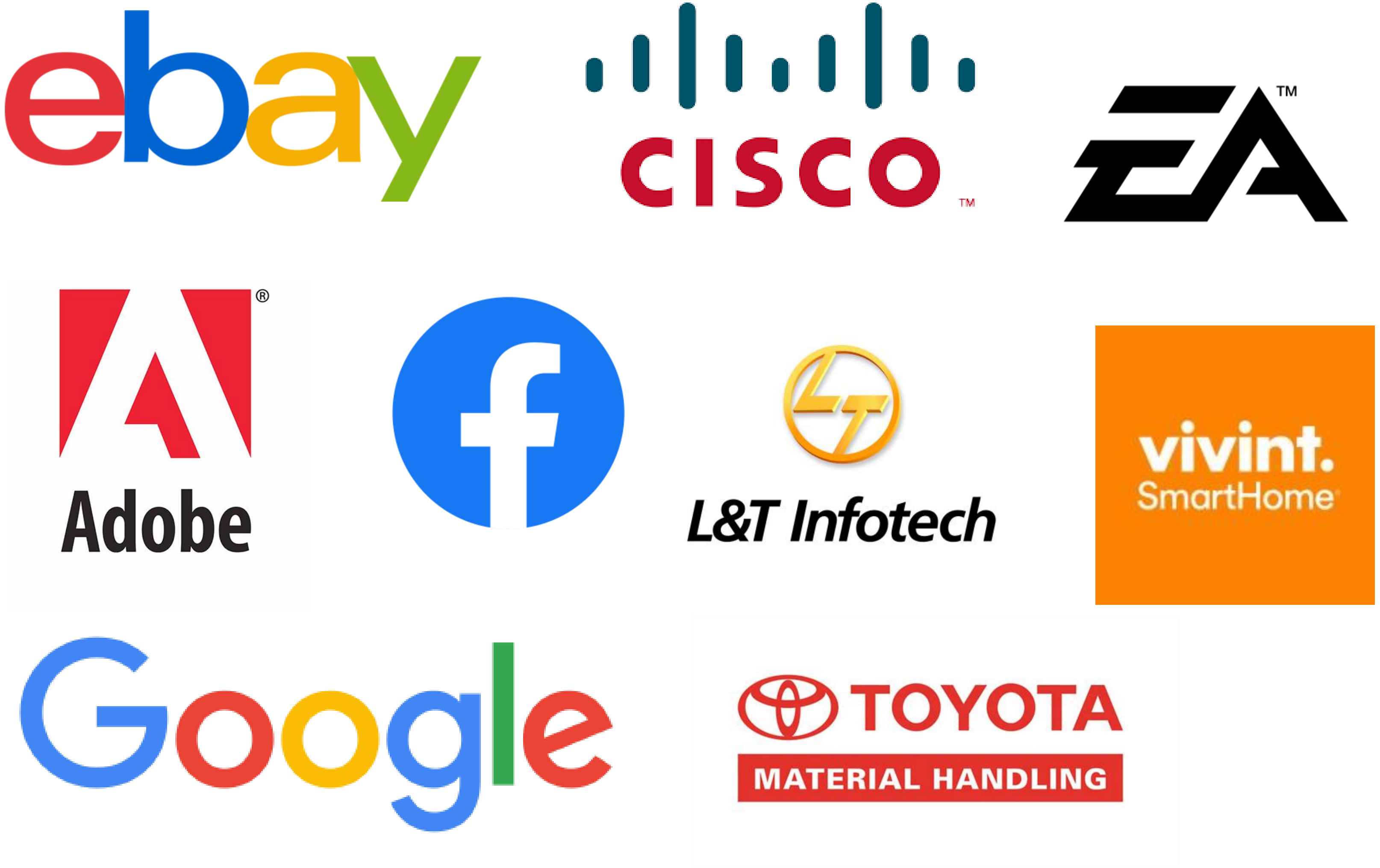
MongoDB Query Language (MQL)

- MongoDB queries are based on JavaScript.
- The language is reasonably easy to learn and many tools are available to query MongoDB data using SQL syntax. Like…
  - ✓ MongoDB BI Connector,
  - ✓ MongoDB Charts,
  - ✓ MongoDB Command Line Interface,
  - ✓ MongoDB Compass,
  - ✓ MongoDB Database Tools, etc.

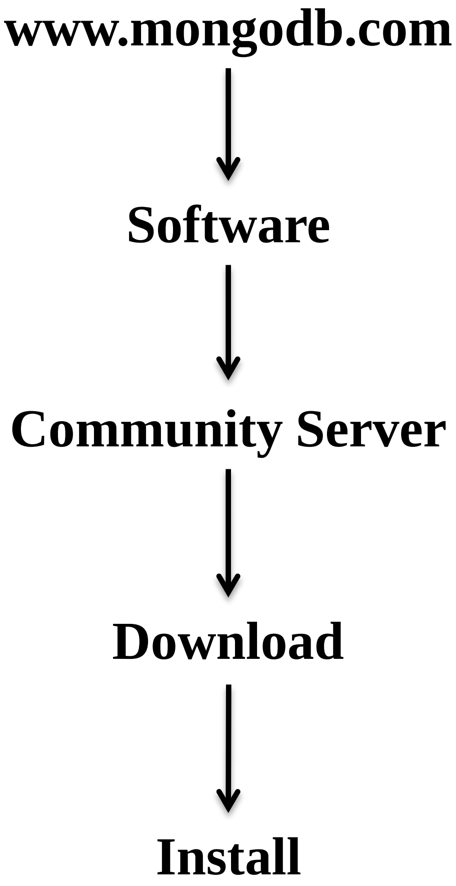
https://docs.mongodb.com/tools/

- When querying data, you have an extraordinary range of options, operators, expressions and filters.

CE: 5.1.4. Few Companies using MongoDB



CE: How to install Mongodb?



www.mongodb.com

Software

Community Server

Download

Install

CE: 5.2.
Data Types and Operators

CE: 5.2.1. Few Data Types supported in MongoDB

- String
- Integer
- Boolean
- Double
- Min/ Max keys - Used to compare a value against the lowest and highest BSON elements.
- Arrays
- Timestamp - This can be handy for recording when a document has been modified or added.
- Object - Used for embedded documents.
- Null - Used to store a Null value.
- Symbol - Used identically to a string; however, it's generally reserved for languages that use a specific symbol type.
- Date - Used to store the current date or time in UNIX time format. You can specify your own date time by creating object of Date and passing day, month, year into it.

CE: 5.2.1. Few Data Types supported in MongoDB (Conti…)

- Object ID - Used to store the document’s ID.
- Binary data - Used to store binary data.
- Code - Used to store JavaScript code into the document.
- Regular expression - Used to store regular expression.

CE: 5.2.2. Relational Operators in MongoDB

Operator	Description
$eq	Equal to
$ne	Not equal to
$gte	Greater than or equal to
$lte	Less than or equal to
$gt	Greater than
$lt	Less than

Exploring NoSQL Query Language

5.3. CRUD Operations

5.4. Built-in Functions and Aggregate Functions

5.5. Import and Export Files

In Practical Tutorial

Another NoSQL database is

Cassandra

CE:
Introduction to Cassandra

CE: What is Cassandra?

- Cassandra is a ***distributed database management system*** designed for handling a high volume of ***structured data*** across commodity servers.

- Cassandra handles the huge amount of data with its distributed architecture.

- Data is placed on different machines with more than one replication factor that provides high availability and no single point of failure.

- Cassandra is managed by Apache.

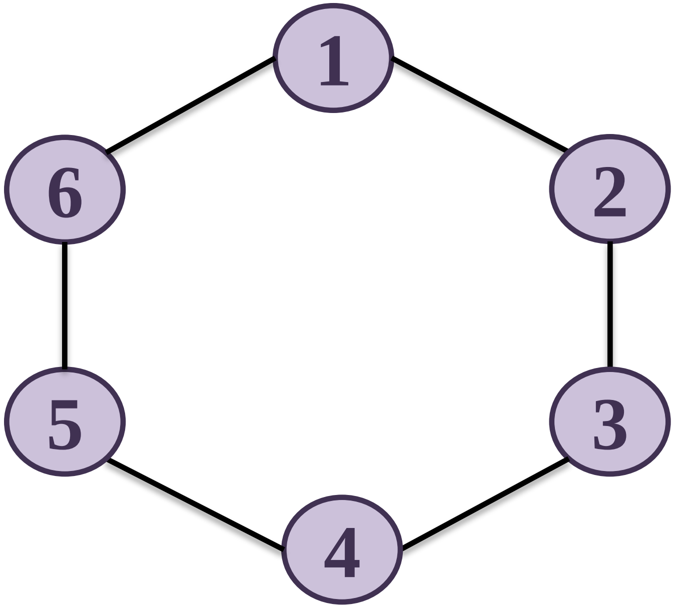
- It was first developed at Facebook for inbox search.

- Facebook open sourced it in July 2008.

- And Apache incubator accepted Cassandra in March 2009.

CE: What is Cassandra? (Conti…)

- It is a top level project of Apache since February 2010.



CE: Features of Apache Cassandra

- Massively Scalable Architecture: All nodes are at the same level which provides operational simplicity and easy scale out.

- Masterless Architecture: Data can be written and read on any node.

- Linear Scale Performance: As more nodes are added, the performance of Cassandra increases.

- No Single point of failure: Cassandra replicates data on different nodes that ensures no single point of failure.

- Fault Detection and Recovery: Failed nodes can easily be restored and recovered.

- Flexible and Dynamic Data Model: Supports datatypes with Fast writes and reads.

CE: Features of Apache Cassandra (Conti…)

- Data Protection: Data is protected with commit log design and build in security like backup and restore mechanisms.

- Tunable Data Consistency: Support for strong data consistency across distributed architecture.

- Multi Data Center Replication: Cassandra provides feature to replicate data across multiple data center.

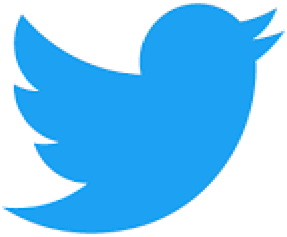
- Data Compression: Cassandra can compress up to 80% data without any overhead.

- Cassandra Query Language: Cassandra provides query language that is similar like SQL language. It makes very easy for relational database developers moving from relational database to Cassandra.

CE: Application of Apache Cassandra

- Messaging
- Internet of things Application
- Product Catalogs and retail apps
  [For shopping cart protection and fast product catalog input and output.
- Social Media Analytics and recommendation engine

CE: Few Companies using Apache Cassandra



CE:
Overview of

CE: Overview of Hive

- Apache Hive is a *data warehouse system* for Hadoop that runs SQL like queries called ***HQL (Hive query language)*** which gets internally converted to map reduce jobs.

- Initially, Hive was developed by Facebook. It supports Data Definition Language, Data Manipulation Language and user defined functions.

- Later the Apache Software Foundation took it up and developed it further as an open source under the name ***Apache Hive***.

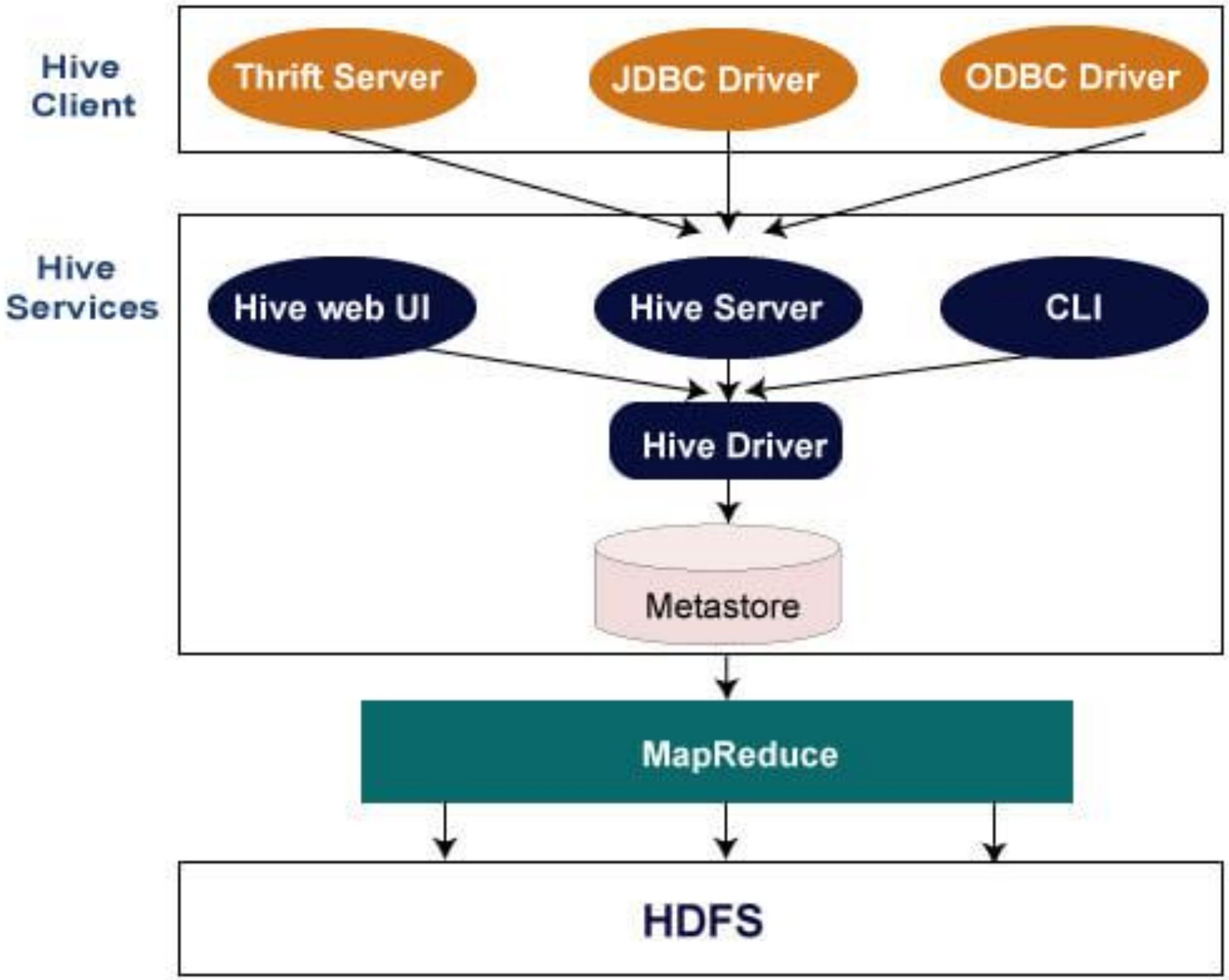
- It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

- It is used by different companies.
  - ✓ For example:
    - Amazon uses it in Amazon Elastic MapReduce.

CE: Overview of Hive (Conti…)

- Hive is not
  - ✓ A relational database
  - ✓ A design for OnLine Transaction Processing (OLTP)
  - ✓ A language for real-time queries and row-level updates

CE: Hive Architecture



CE: Features of Hive

- Hive is fast and scalable.
- It provides SQL type language for querying called HiveQL or HQL, that are implicitly transformed to MapReduce or Spark jobs.
- It stores schema in a database and processed data into HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It is designed for OLAP.
- It supports SQL filters, group-by and order-by clauses.
- It can operate on compressed data stored in the Hadoop ecosystem.

CE: Features of Hive (Conti…)

- HQL is easy to code.
- It supports user-defined functions (UDFs), where user can provide its functionality.

CE: Limitations of Hive

- Hive is not capable of handling real-time data.
- It is not designed for online transaction processing.
- Hive queries contain high latency.

CE:
Hive Data Types and File format

CE: Hive Data Types

❖ **Primitive Data Types:**

- **Numeric:**
  - ✓ Tinyint
  - ✓ Smallint
  - ✓ Int
  - ✓ Bigint
  - ✓ Float
  - ✓ Double
- **String:**
  - ✓ String
  - ✓ Varchar
  - ✓ Char
- **Miscellaneous:**
  - ✓ Boolean
  - ✓ Binary

❖ **Collection Data Types;**

- **Struct**
- **Map**
- **Array**

CE: Hive File format

- Text File
- Sequential File
- RCFile (Record Columnar File)

CE:
Hive Query Language (HQL)

CE: Hive Query Language (HQL)

- DDL
- DML

CE: Differences between Hive and Pig

Hive	Pig
1. Hive is commonly used by Data Analysts.	1. Pig is commonly used by programmers.
2. It follows SQL-like queries.	2. It follows the data-flow language.
3. It can handle structured data.	3. It can handle semi-structured data.
4. It works on server-side of HDFS cluster.	4. It works on client-side of HDFS cluster.
5. Hive is slower than Pig.	5. Pig is comparatively faster than Hive.

Thank you