# Syllabus

- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.
- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.
- An Analytics Project-
  - Communicating,
  - operationalizing,
  - creating final deliverables.

# Syllabus

- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.

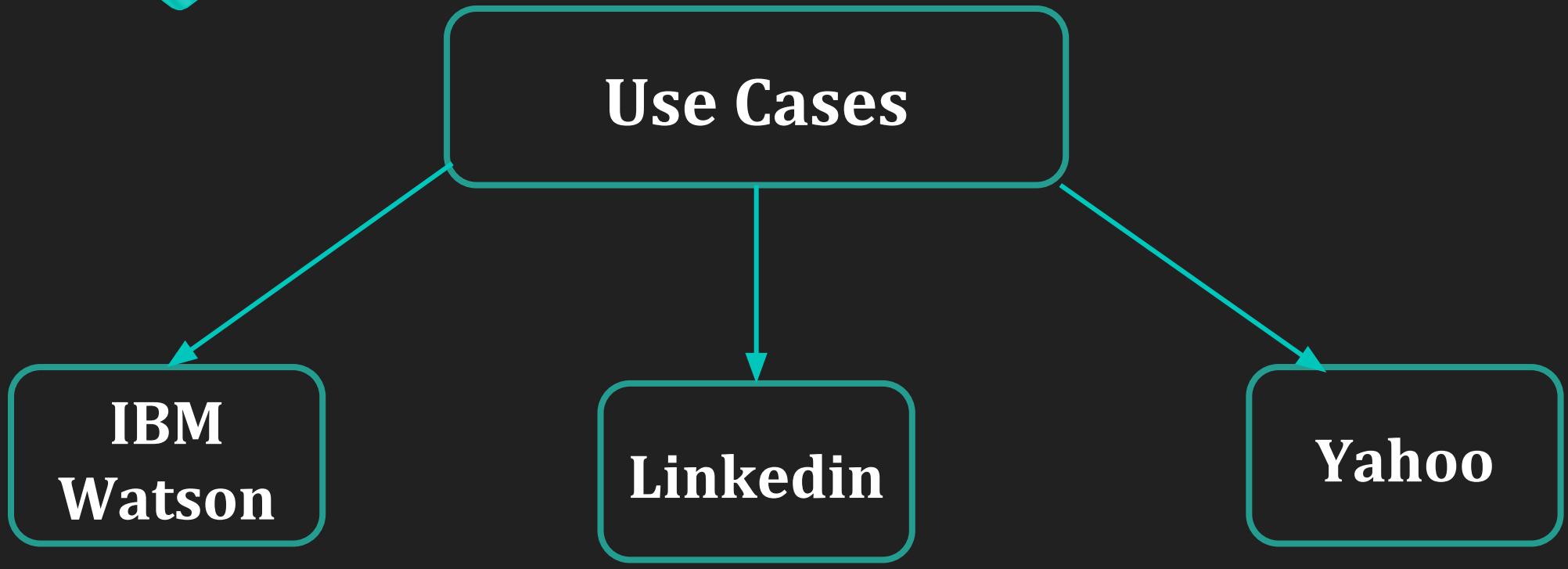
# Analytics for Unstructured Data

- Structured: A specific and consistent format (for example, a data table)
- Semi-structured: A self-describing format (for example, an XML file)
- Quasi-structured: A somewhat inconsistent format (for example, a hyperlink)
- Unstructured: An inconsistent format (for example, video)

# Analytics for Unstructured Data

- Use cases for MapReduce.
- The MapReduce paradigm offers the means to break a large task into smaller tasks, run tasks in parallel, and consolidate the outputs of the individual tasks into the final output.
- Apache Hadoop includes a software implementation of MapReduce.

# Analytics for Unstructured Data



# Analytics for Unstructured Data

## IBM Watson

- In 2011, IBM's computer system **Watson participated in the U.S. television game show Jeopardy** against two of the best Jeopardy champions in the show's history.
- In the game, the contestants are provided a clue such as "He likes his martinis shaken, not stirred" and the correct response, phrased in the form of a question, would be, "Who is James Bond?"
- Over the **three-day tournament,** Watson was able to defeat the two human contestants.
- **To educate Watson**, Hadoop was utilized to process various data sources such as encyclopedias, dictionaries, news wire feeds, literature, and the entire contents of Wikipedia.

# Analytics for Unstructured Data

## IBM Watson

- For each clue provided during the game, **Watson had to perform the following tasks in less than 3 seconds**
  - Deconstruct the provided clue into words and phrases
  - Establish the grammatical relationship between the words and the phrases
  - Create a set of similar terms to use in Watson's search for a response
  - Use Hadoop to coordinate the search for a response across terabytes of data
  - Determine possible responses and assign their likelihood of being correct
  - Actuate the buzzer Provide a syntactically correct response in English
- Watson is being used in the **medical profession to diagnose patients** and provide treatment recommendations

# Analytics for Unstructured Data

## LinkedIn

- LinkedIn is an online professional network of 250 million users in 200 countries as of early 2014 .
- LinkedIn provides several free and subscription-based services, such as **company information pages, job postings, talent searches, social graphs of one's contacts, personally tailored news feeds**, and access to discussion groups, including a Hadoop users group

# Analytics for Unstructured Data

**LinkedIn**

- **LinkedIn utilizes Hadoop for the following purposes :**
  - Process daily production database transaction logs
  - Examine the users' activities such as views and clicks
  - Feed the extracted data back to the production systems
  - Restructure the data to add to an analytical database
  - Develop and test analytical models

# Analytics for Unstructured Data

**Yahoo!**

- As of 2012, Yahoo! has one of the largest publicly announced **Hadoop deployments at 42,000 nodes across several clusters utilizing 350 petabytes of raw storage .**

# Analytics for Unstructured Data

**Yahoo!**

- **Yahoo!'s Hadoop applications** include the following :
  - Search index creation and maintenance
  - Web page content optimization
  - Web ad placement optimization
  - Spam filters
  - Ad-hoc analysis and analytic model development
- Prior to deploying Hadoop, it took **26 days to process three years' worth of log data**. With **Hadoop**, the processing time was **reduced to 20 minutes.**

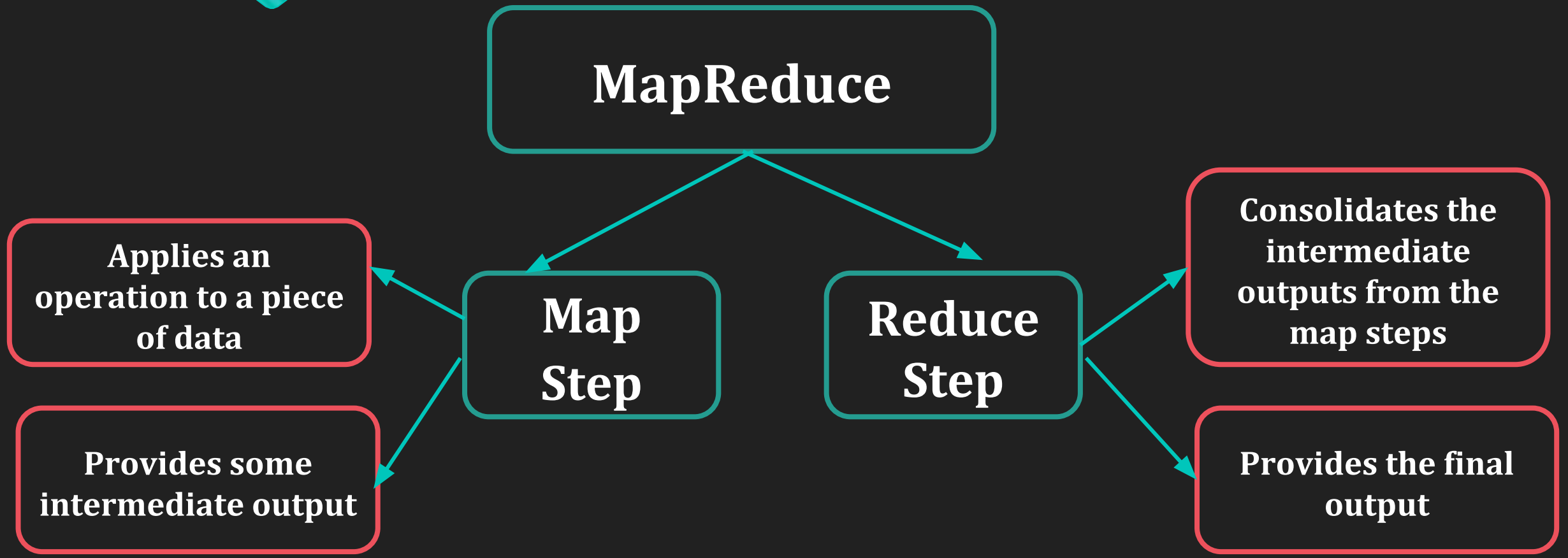
# Syllabus

- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.
-

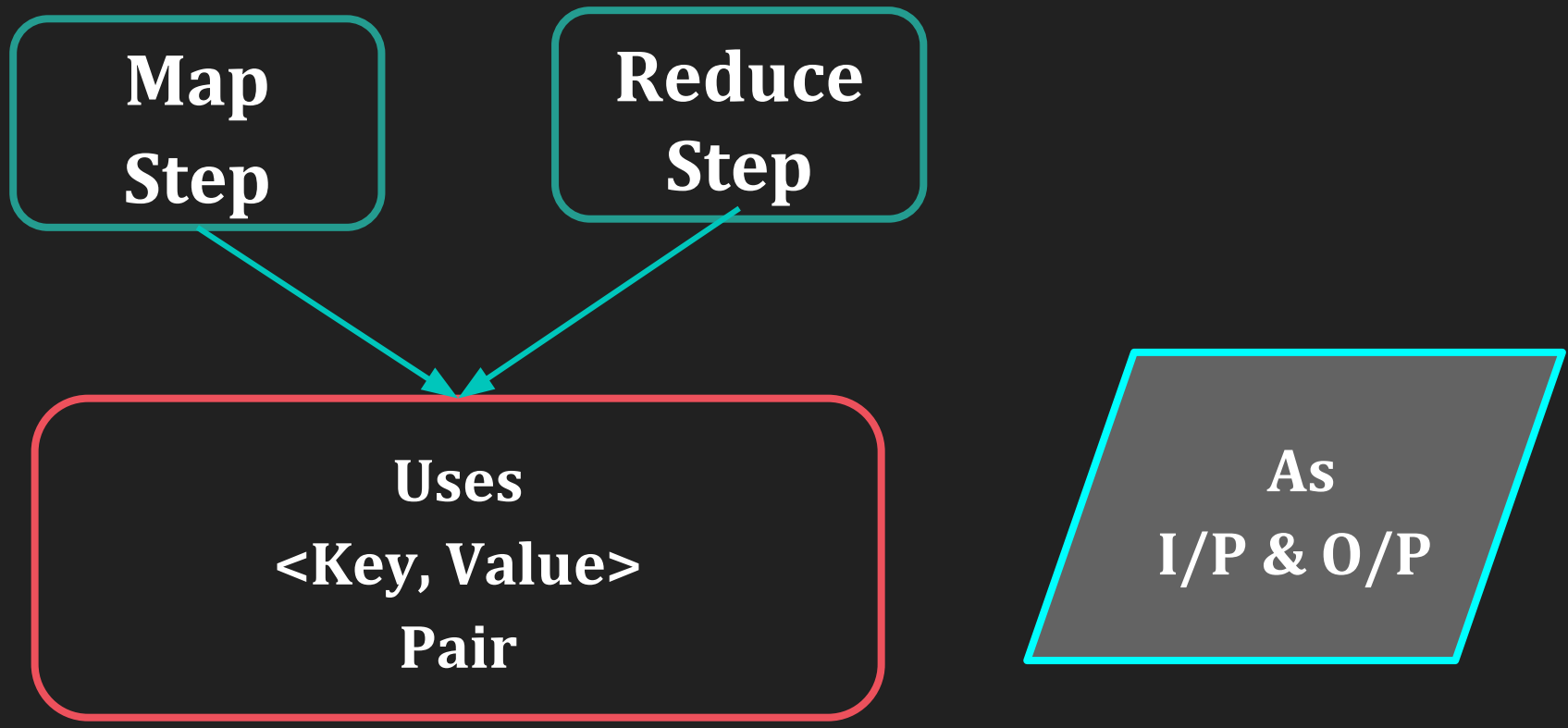
# MapReduce

- MapReduce paradigm provides the means to
  - Break a large task into smaller tasks,
  - Run the tasks in parallel, and
  - Consolidate the outputs of the individual tasks into the final output

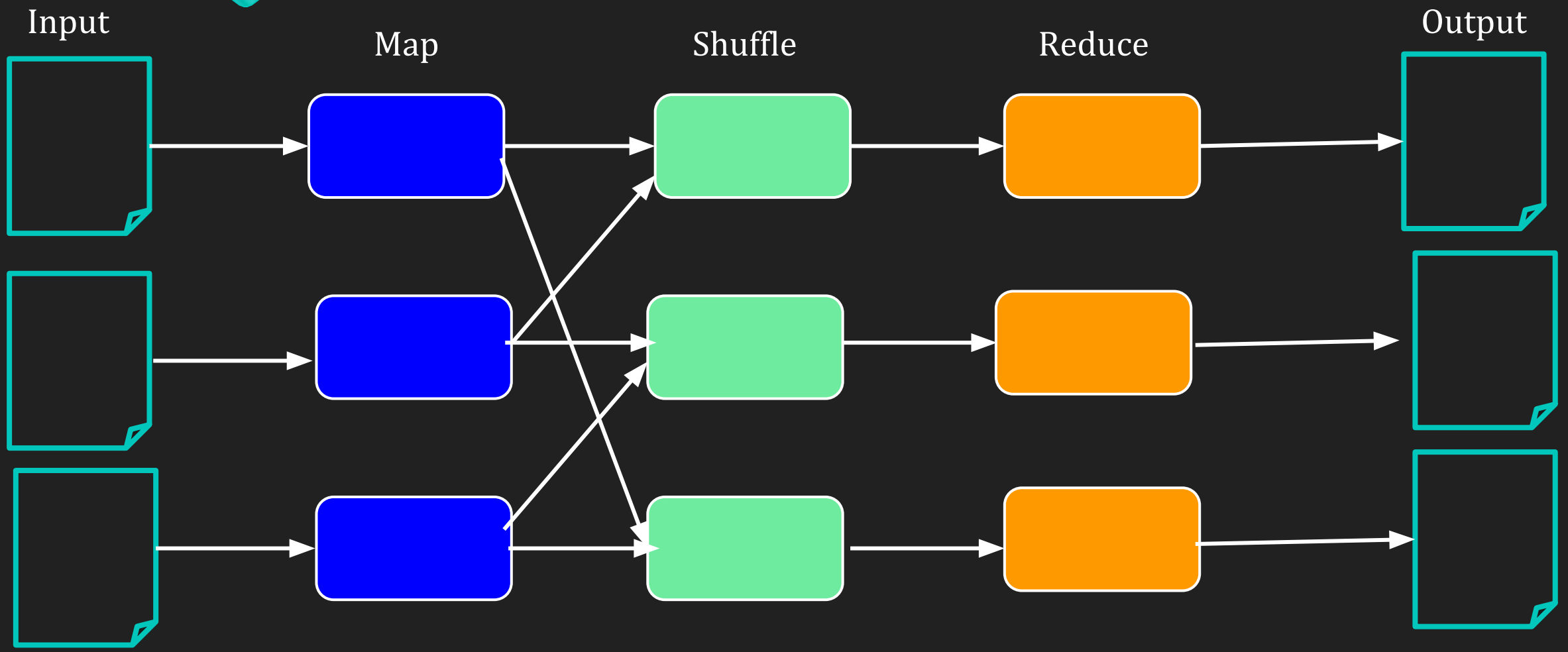
# MapReduce


MapReduce
Map Step
Reduce Step
Applies an operation to a piece of data
Provides some intermediate output
Consolidates the intermediate outputs from the map steps
Provides the final output

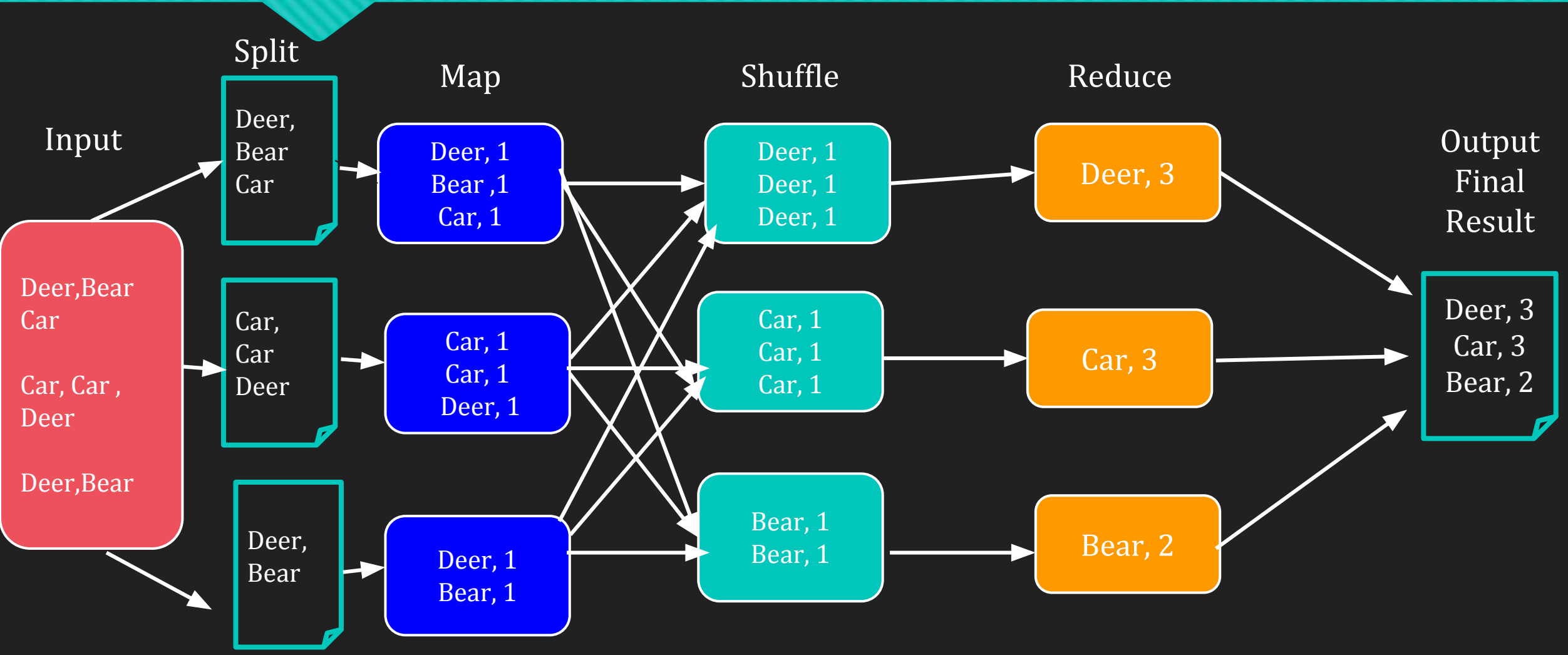
# MapReduce



# MApReduce Example- Word Count



# MApReduce Example- Word Count



# Syllabus

- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.

# Apache Hadoop

- Hadoop Overview
- Hadoop Distributed File System (HDFS)
- Structuring a MapReduce Job in Hadoop
- Additional Considerations in Structuring a MapReduce Job
- Developing and Executing a Hadoop MapReduce Program
- Yet Another Resource Negotiator (YARN)

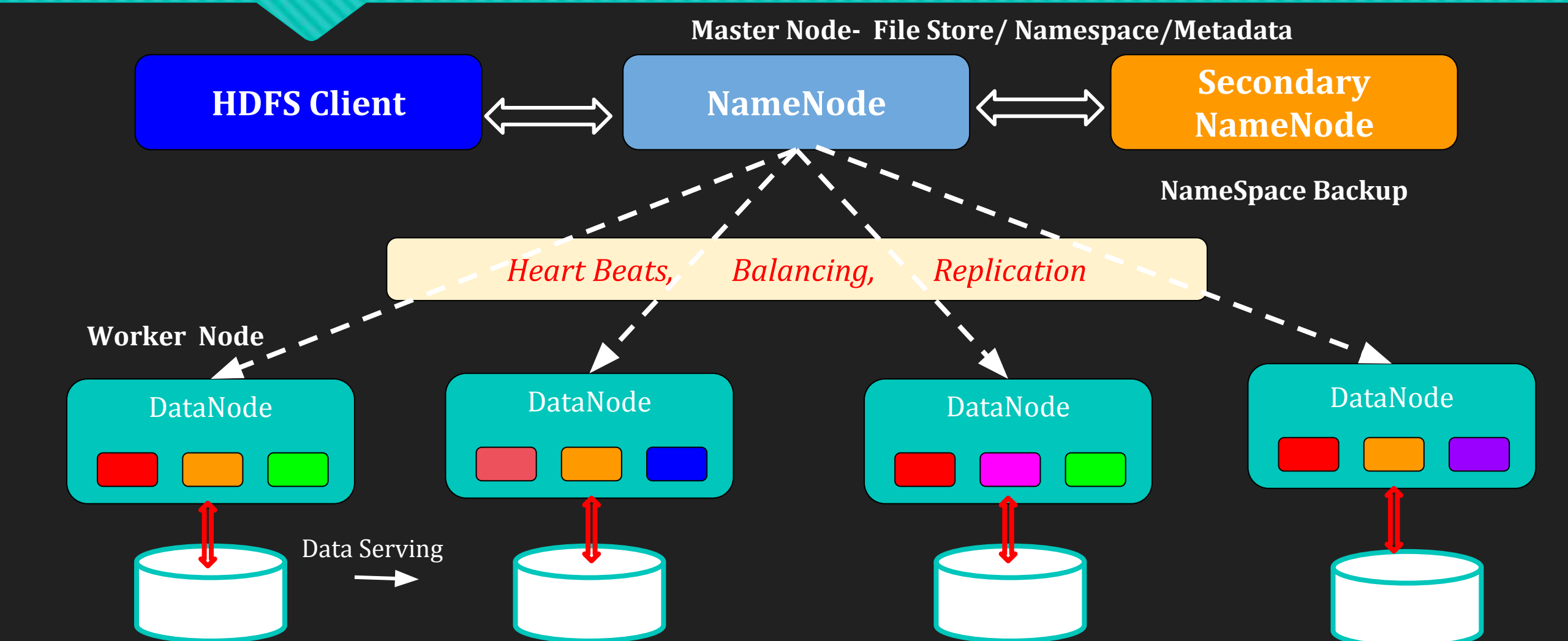
# Hadoop Overview

- Hadoop is an open source framework, from the Apache foundation,
- Capable of processing large amounts of heterogeneous data sets in a distributed fashion across clusters of commodity computers and hardware using a simplified programming model.
- Hadoop provides a reliable shared storage and analysis system.

# Hadoop Distributed File System (HDFS)

- Hadoop Distributed File System (HDFS) is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. HDFS was designed to be a scalable, fault-tolerant, distributed storage system that works closely with MapReduce.
- HDFS has a master/slave architecture. An HDFS cluster consists of a single Name Node (a master server that manages the file system namespace and regulates access to files by clients) and a number of Data Nodes. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks (typically 64Mb in size) and these blocks are stored in a set of Data Nodes.
- The Name Node executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data Nodes. The Data Nodes are responsible for serving read and write requests from the file system's clients. The Data Nodes also perform block creation, deletion, and replication upon instruction from the Name Node.

# Hadoop Distributed File System (HDFS)



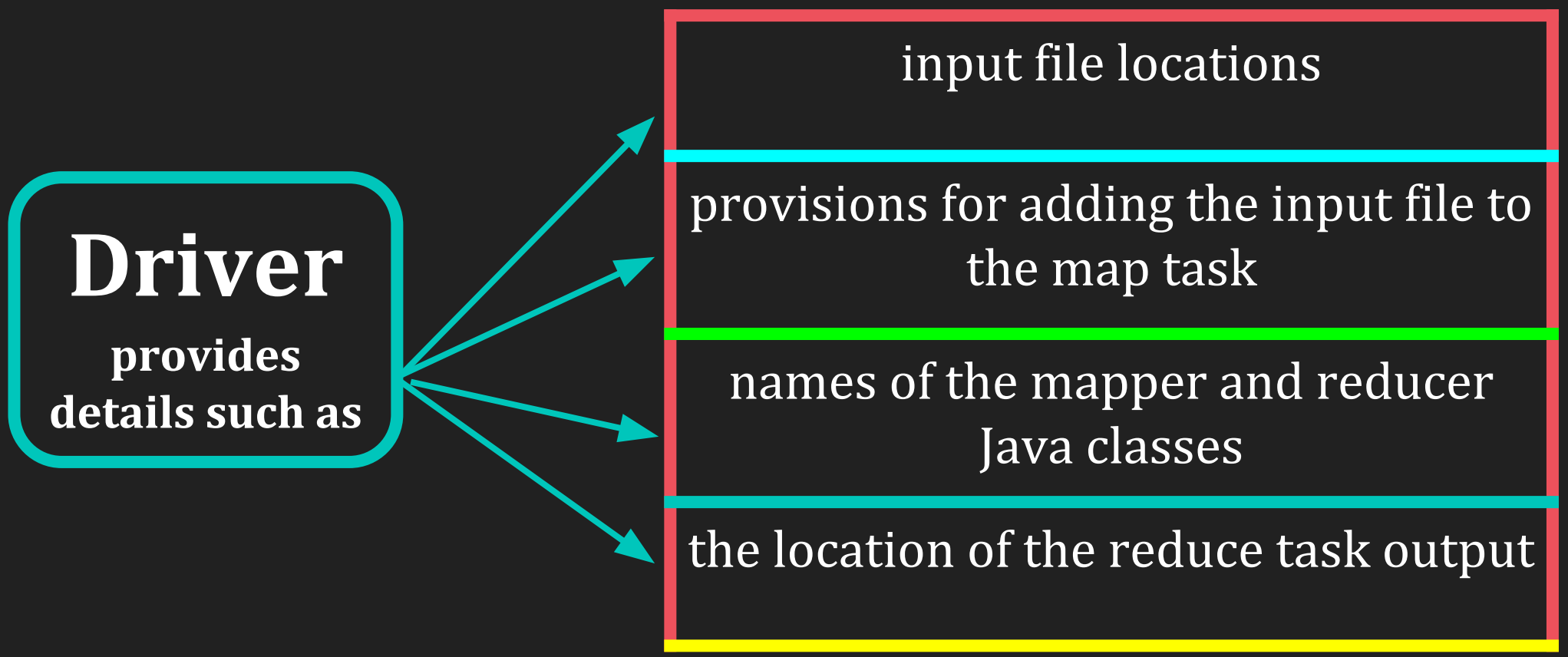
# Structuring a MapReduce Job in Hadoop

**How a MapReduce job is run in Hadoop?**

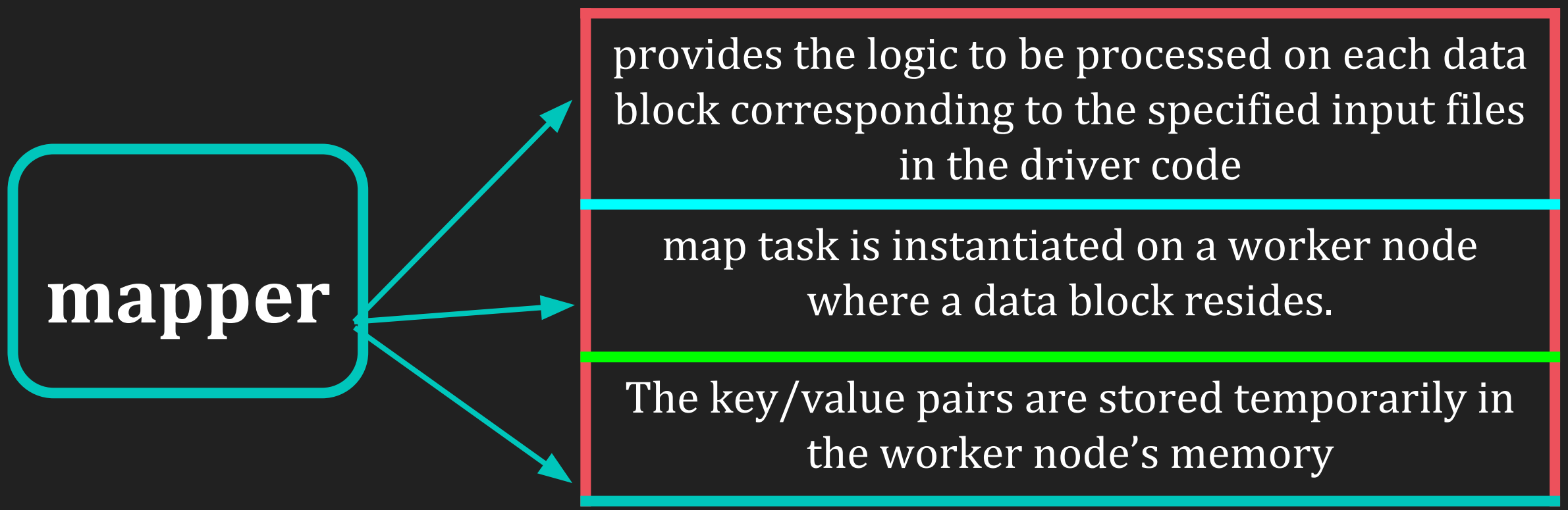
**A typical MapReduce program in Java consists of three classes**

- Driver
- Mapper
- Reducer

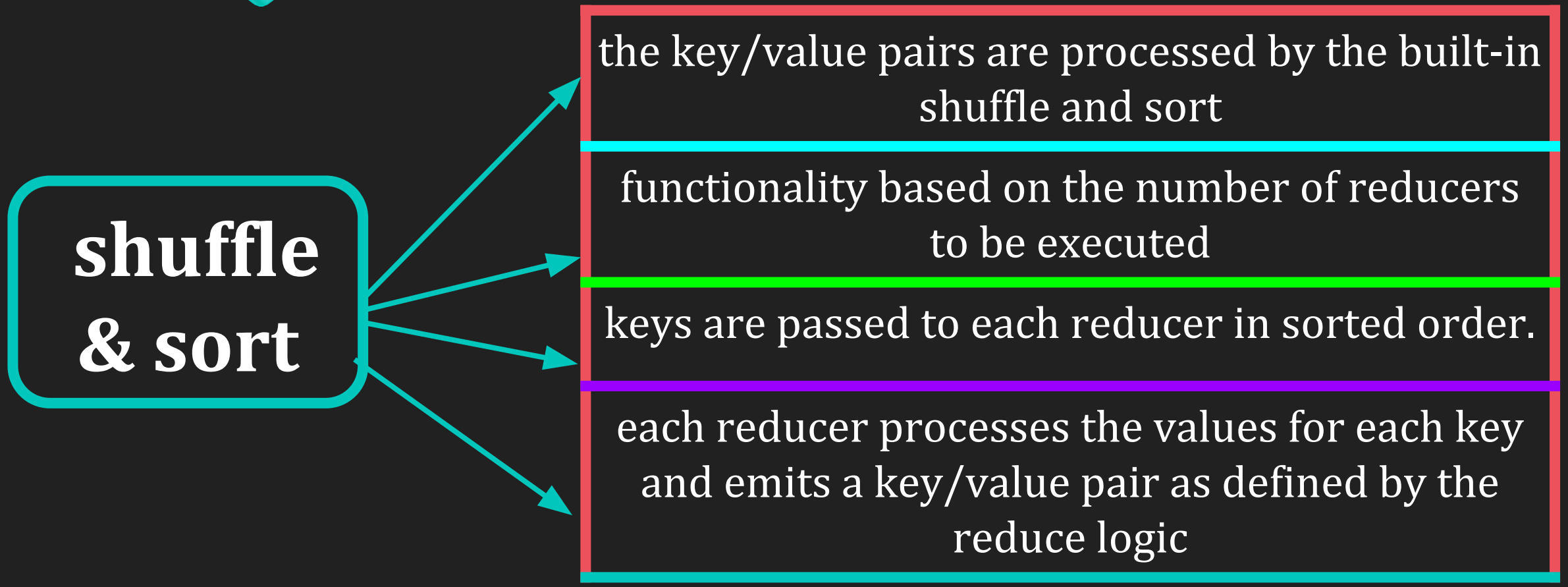
# Structuring a MapReduce Job in Hadoop



# Structuring a MapReduce Job in Hadoop



# Structuring a MapReduce Job in Hadoop



# Additional Considerations in Structuring a MapReduce Job

**Several Hadoop features provide additional functionality to a MapReduce job.**

## Combiner

Apply between map task & shuffle & sort

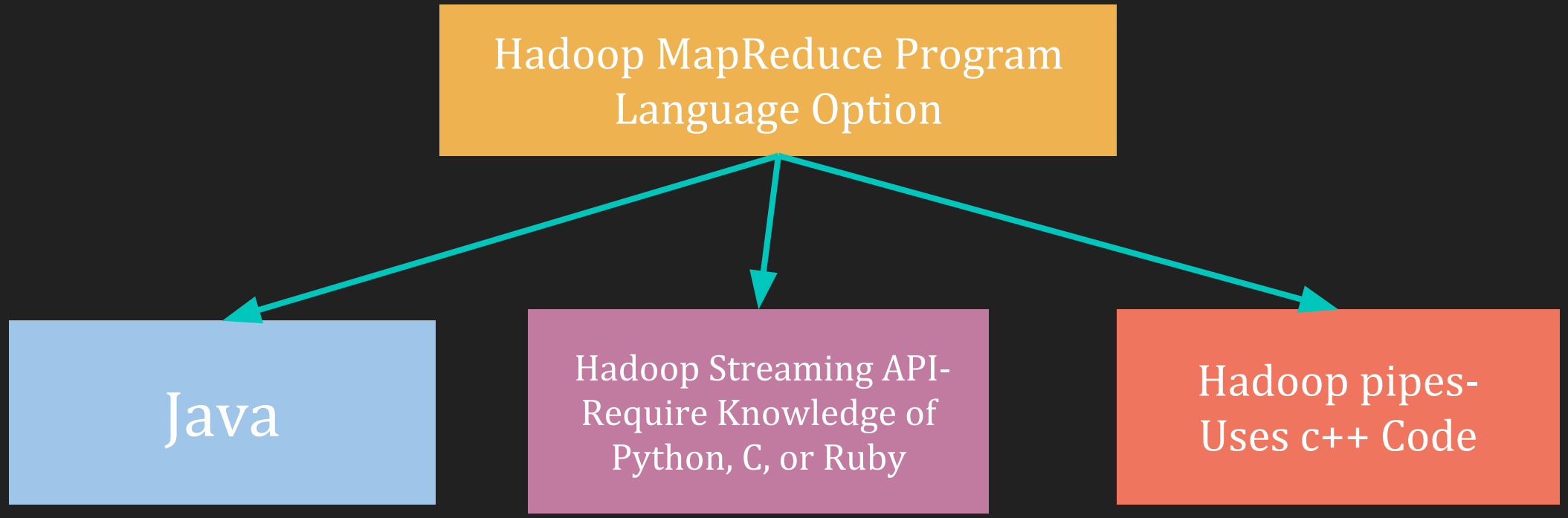
minimizes the amount of intermediate map output

## partitioner

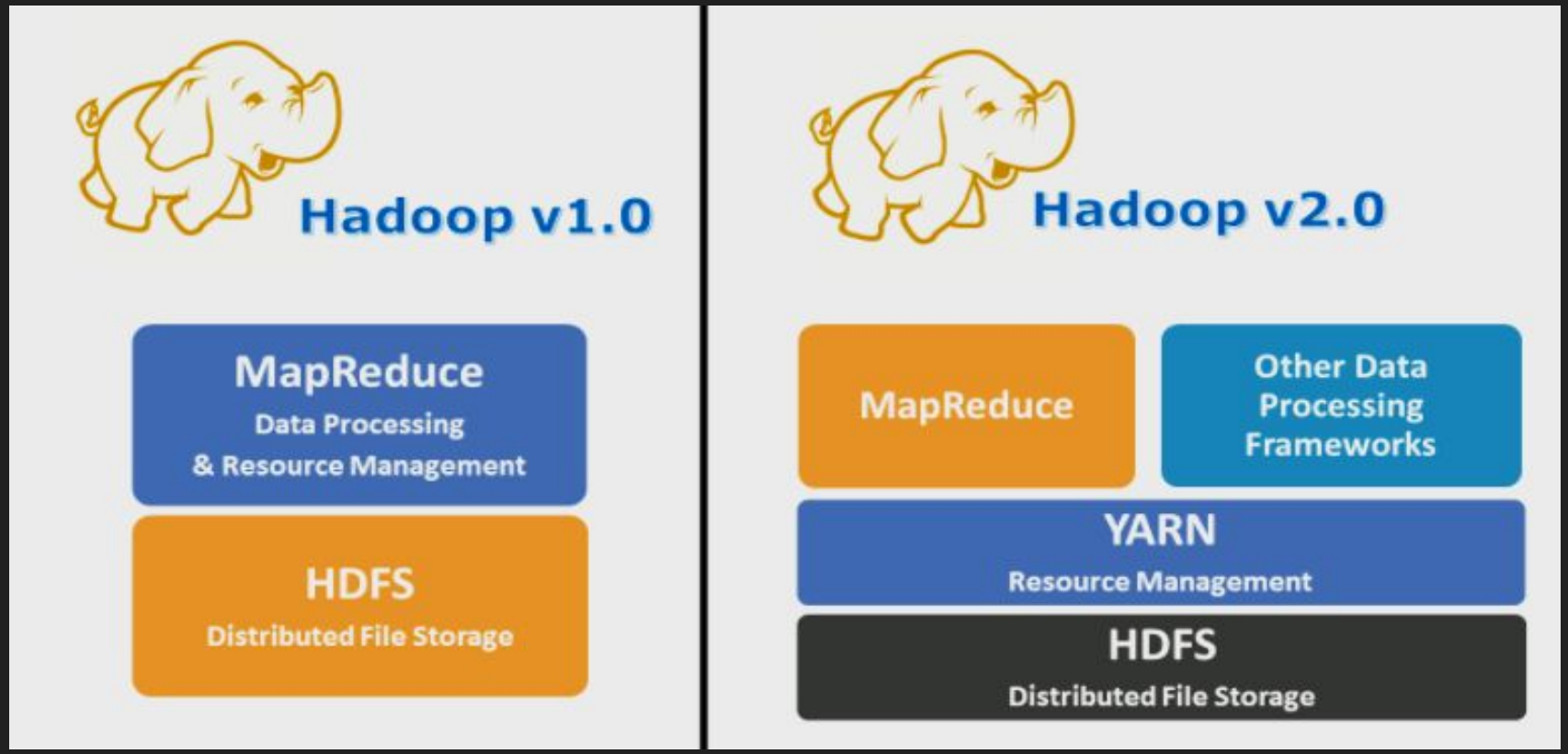
separate the output into separate files for for subsequent analysis.

ensure that the workload is evenly distributed across the reducers

# Developing and Executing a Hadoop MapReduce Program



# Yet Another Resource Negotiator (YARN)



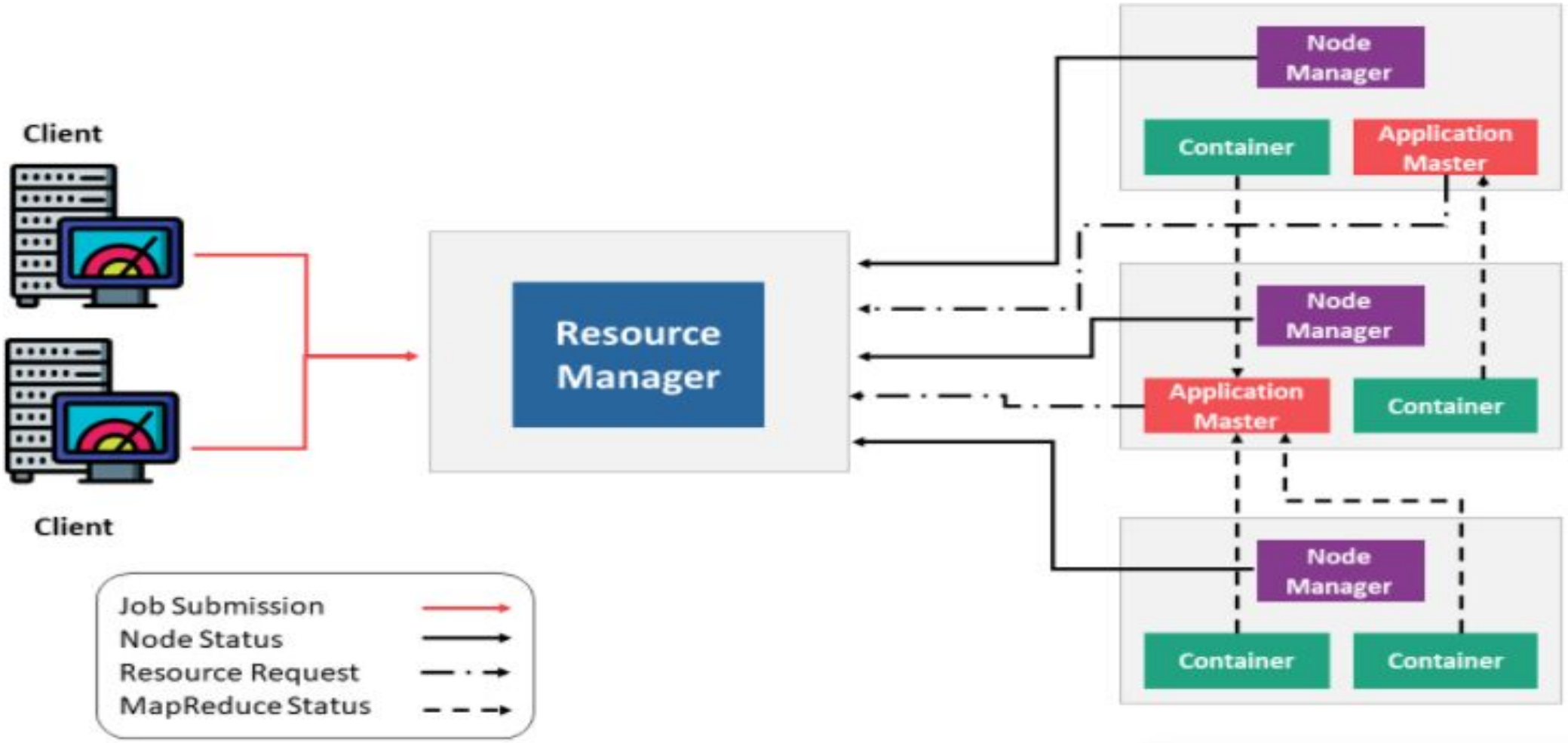

https://www.edureka.co/blog/hadoop-yarn-tutorial/

# Yet Another Resource Negotiator (YARN)

- The technology used for job scheduling and resource management and one of the main components in Hadoop is called Yarn.
- Yarn stands for Yet Another Resource Negotiator though it is called as Yarn by the developers.
- Yarn was previously called MapReduce2 and Nextgen MapReduce.
- This enables Hadoop to support different processing types.
- It runs interactive queries, streaming data and real time applications.
- Also it supports broader range of different applications.
- Yarn combines central resource manager with different containers.
- It can combine the resources dynamically to different applications and the operations are monitored well.

# Yet Another Resource Negotiator (YARN)

https://www.edureka.co/blog/hadoop-yarn-tutorial/



# Yet Another Resource Negotiator (YARN)

**Resource Manager:** Runs on a master daemon and manages the resource allocation in the cluster.

**Node Manager:** They run on the slave daemons and are responsible for the execution of a task on every single Data Node.

**Application Master**: Manages the user job lifecycle and resource needs of individual applications. It works along with the Node Manager and monitors the execution of tasks.

**Container**: Package of resources including RAM, CPU, Network, HDD etc on a single node.

# Syllabus

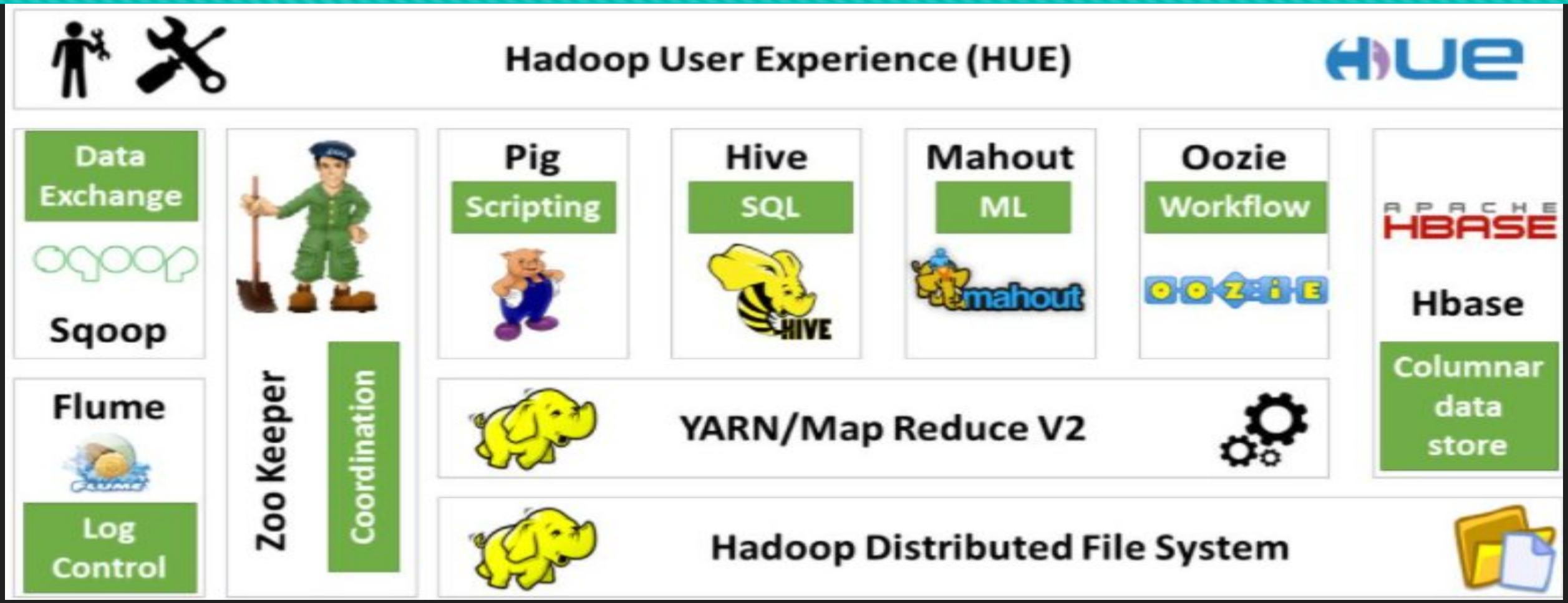
- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.
- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.
- An Analytics Project-
  - Communicating,
  - operationalizing,
  - creating final deliverables.

# Syllabus

- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

# The Hadoop Ecosystem

collection of technologies/framework that work in conjunction with Hadoop

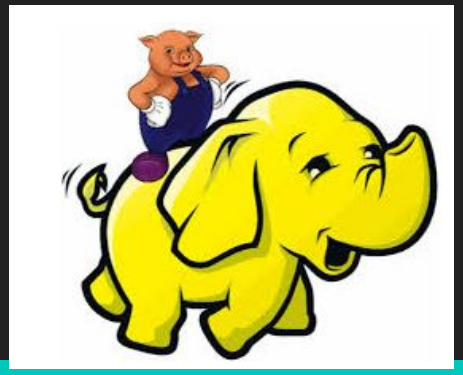



https://www.ktexperts.com/hadoop-ecosystem/

# Syllabus

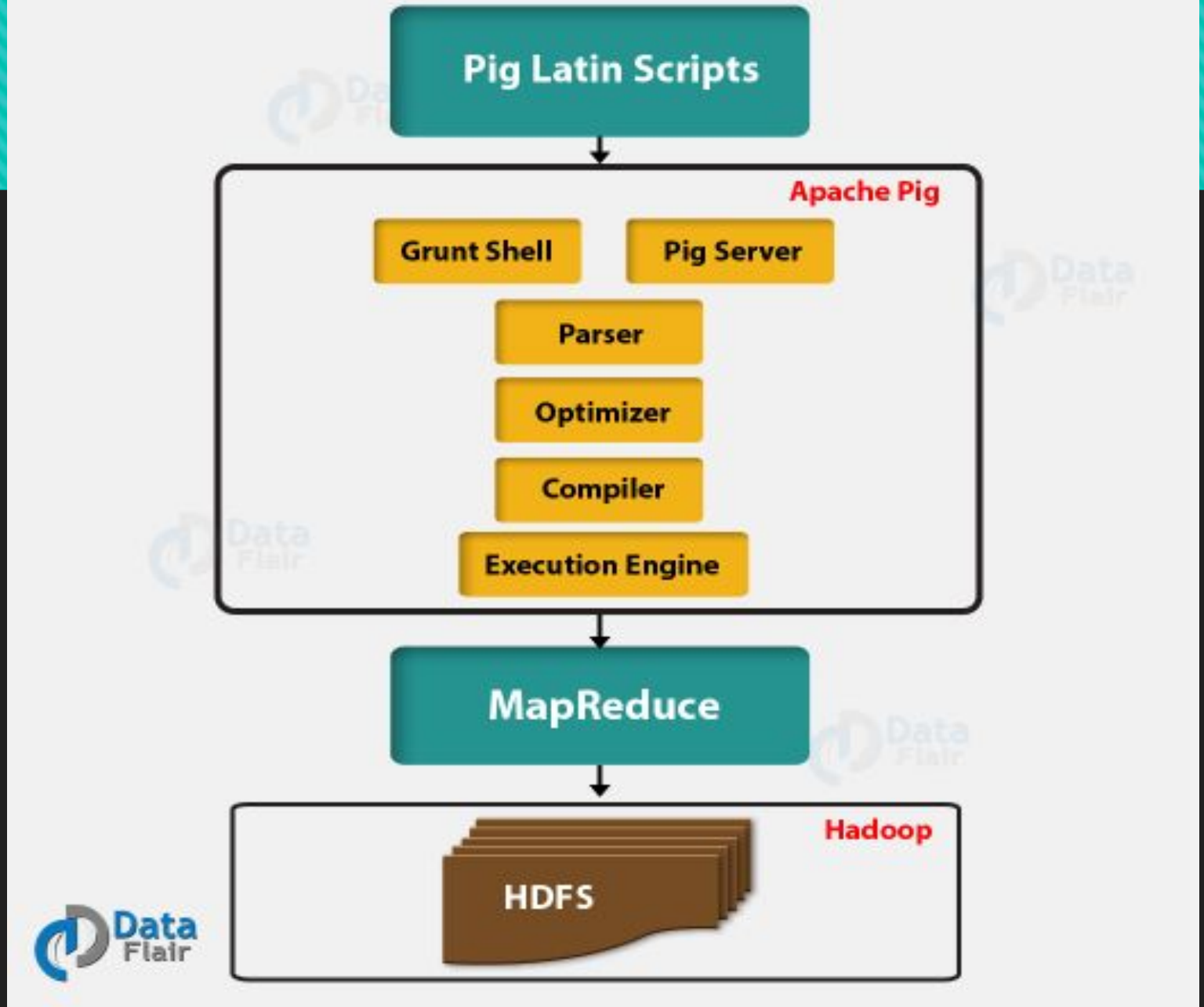
- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

# Pig



# Pig Architecture

https://data-flair.training/blogs/pig-architecture/



# Pig

- Apache Pig consists of a
  - data flow language,
  - Pig Latin, and
  - environment to execute the Pig code.

- The main benefit of using Pig is to utilize the power of MapReduce in a distributed system,
- while simplifying the tasks of developing and executing a MapReduce job.

# Pig

- Pig include entering the Pig execution environment by typing pig at the command prompt and then entering a sequence of Pig instruction lines at the grunt prompt.

**Example :**

$ pig grunt> records = LOAD ‘/user/customer.txt’ AS (cust_id:INT,

first_name:CHARARRAY, last_name:CHARARRAY, email_address:CHARARRAY);

grunt> filtered_records = FILTER records BY email_address matches ‘.*@isp.com’;

grunt> STORE filtered_records INTO ‘/user/isp_customers’;

grunt> quit

# Pig- Build in Functions

- Eval
- Load/Store
- Math
- String
- DateTime

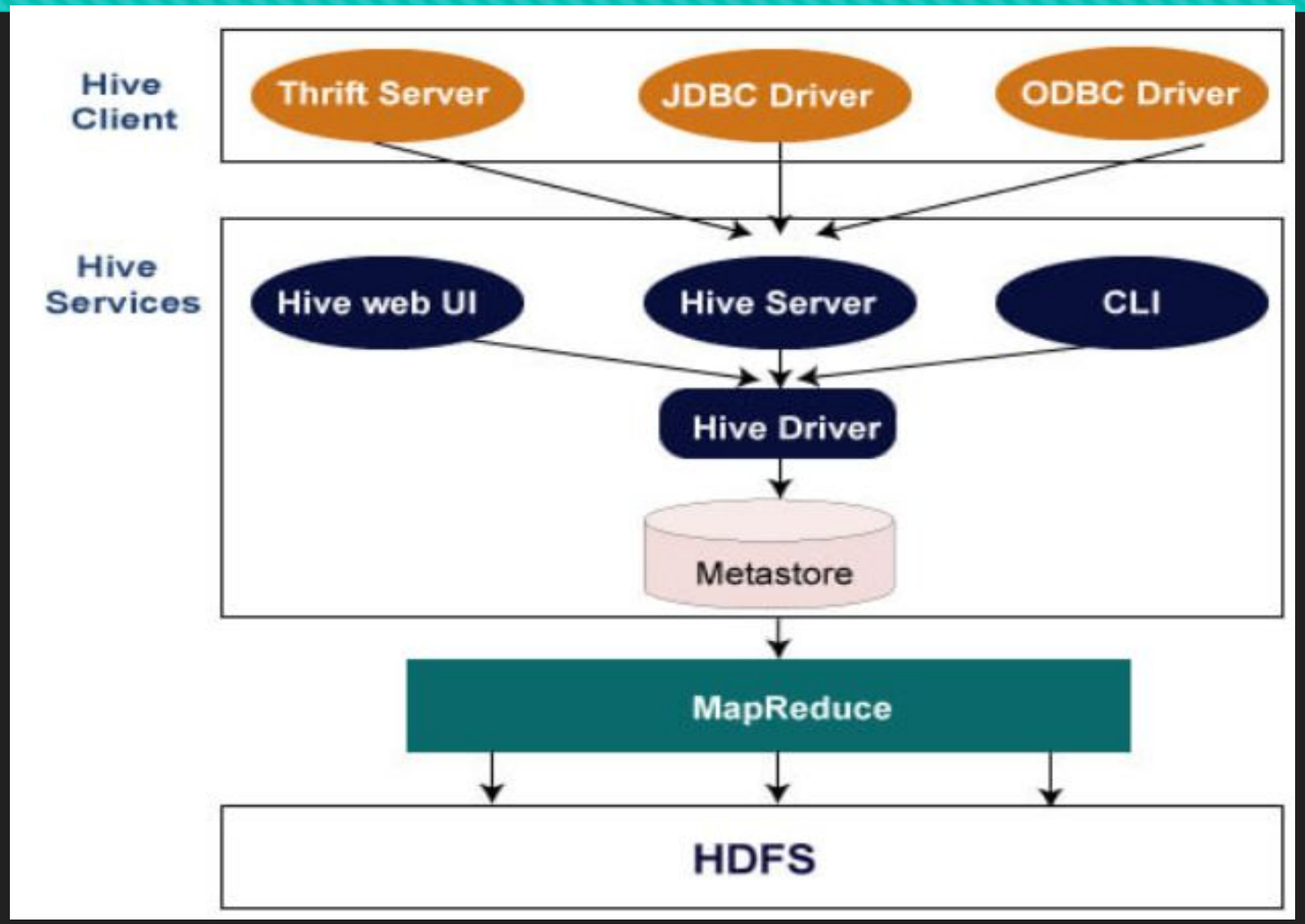
# Syllabus

- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

https://www.javatpoint.com/hive-architecture

# Hive


## Hive Architecture



# Hive

- Apache Hive enables users to process data without explicitly writing MapReduce code.
- Hive language, HiveQL (Hive Query Language), resembles Structured Query Language (SQL)
- A Hive table structure consists of rows and columns.
- The rows typically correspond to some record, transaction, or particular entity (for example, customer) detail.
- The values of the corresponding columns represent the various attributes or characteristics for each row.
- Additionally, a user may consider using Hive if the user has experience with SQL and the data is already in HDFS.
- Hive is not intended for real-time querying

# Hive

When to use Hive?

- Data easily fits into a table structure.
- Data is already in HDFS. (Note: Non-HDFS files can be loaded into a Hive table.)
- Developers are comfortable with SQL programming and queries.
- There is a desire to partition datasets based on time. (For example, daily updates are added to the Hive table.)
- Batch processing is acceptable.

# Hive

- To start hive simply type hive on command prompt.
- $ hive
- hive>
- From this environment, a user can define new tables, query them, or summarize their contents.
- **hive> create table customer ( cust_id bigint, first_name string, last_name string, email_address string) row format delimited fields terminated by '\t';**
- ('\t')-delimited HDFS file

# Hive

- To load the customer table with the contents of HDFS file, customer.txt
- **hive> load data inpath '/user/customer.txt' into table customer;**

- HiveQL query is executed to count the number of records in the newly created table, customer.
- **hive> select count(*) from customer;**

# Following are some Hive use cases:

- Exploratory or ad-hoc analysis of HDFS data: Data can be queried, transformed, and exported to analytical tools, such as R.
- Extracts or data feeds to reporting systems, dashboards, or data repositories such as HBase: Hive queries can be scheduled to provide such periodic feeds.
- Combining external structured data to data already residing in HDFS: Hadoop is excellent for processing unstructured data, but often there is structured data residing in an RDBMS, such as Oracle or SQL Server, that needs to be joined with the data residing in HDFS. The data from an RDBMS can be periodically added to Hive tables for querying with existing data in HDFS.

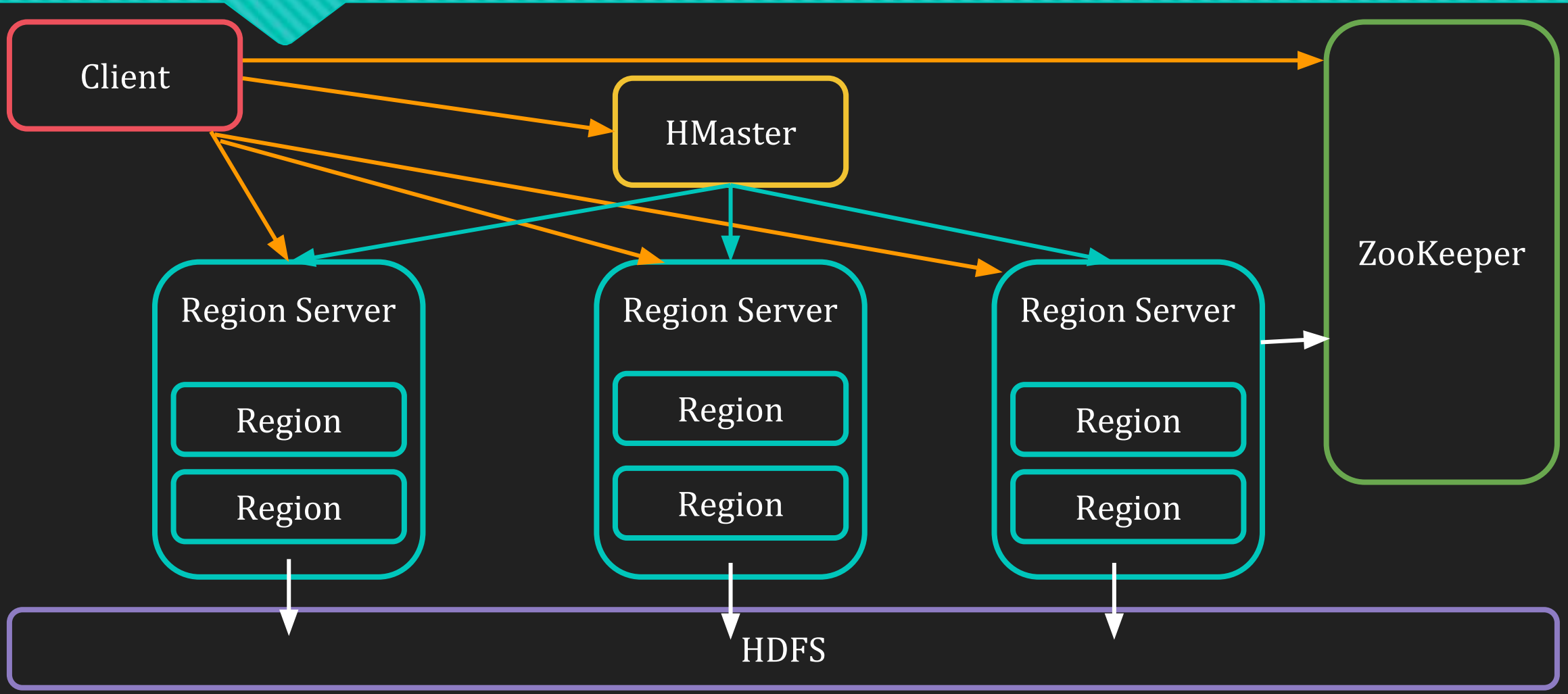
# Syllabus

- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

# HBase

- Apache HBase is capable of providing real-time read and write access to datasets with billions of rows and millions of columns.
- The HBase design is based on Google's 2006 paper on Bigtable.
- In 2010, Facebook began to use HBase for its user messaging infrastructure, which accommodated 350 million users sending 15 billion messages per month

# HBase Architecture and Data Model



# HBase- HBase Components

- **HMaster -**It acts as a monitoring agent to monitor all Region Server instances present in the cluster and acts as an interface for all the metadata changes.
- **HRegionserver**- When Region Server receives writes and read requests from the client, it assigns the request to a specific region, where the actual column family resides.
- **HRegions**- HRegions are the basic building elements of HBase cluster that consists of the distribution of tables and are comprised of Column families
- **Zookeeper** - In HBase, Zookeeper is a centralized monitoring server which maintains configuration information and provides distributed synchronization.
- **HDFS**-HDFS is a Hadoop distributed file system, as the name implies it provides a distributed environment for the storage and it is a file system designed in a way to run on commodity hardware.

# HBase

- HBase is built on top of HDFS. HBase uses a key/value structure to store the contents of an HBase table. Each value is the data to be stored at the intersection of the row, column, and version.
- Each key consists of the following elements.
  - Row length
  - Row (sometimes called the row key)
  - Column family length
  - Column family
  - Column qualifier
  - Version
  - Key type

# Use Cases for HBase- Facebook

- A use case is the storage and search access of messages. In 2010, Facebook implemented such a system using HBase. At the time, Facebook's system was handling more than 15 billion user-to-user messages per month and 120 billion chat messages per month.
- Using each word in each user's message, an HBase table was designed as follows:
  - The row was defined to be the user ID.
  - The column qualifier was set to a word that appears in the message.
  - The version was the message ID.
  - The cell's content was the offset of the word in the message.
- This implementation allowed Facebook to provide auto-complete capability in the search box and to return the results of the query quickly, with the most recent messages at the top.

# Syllabus

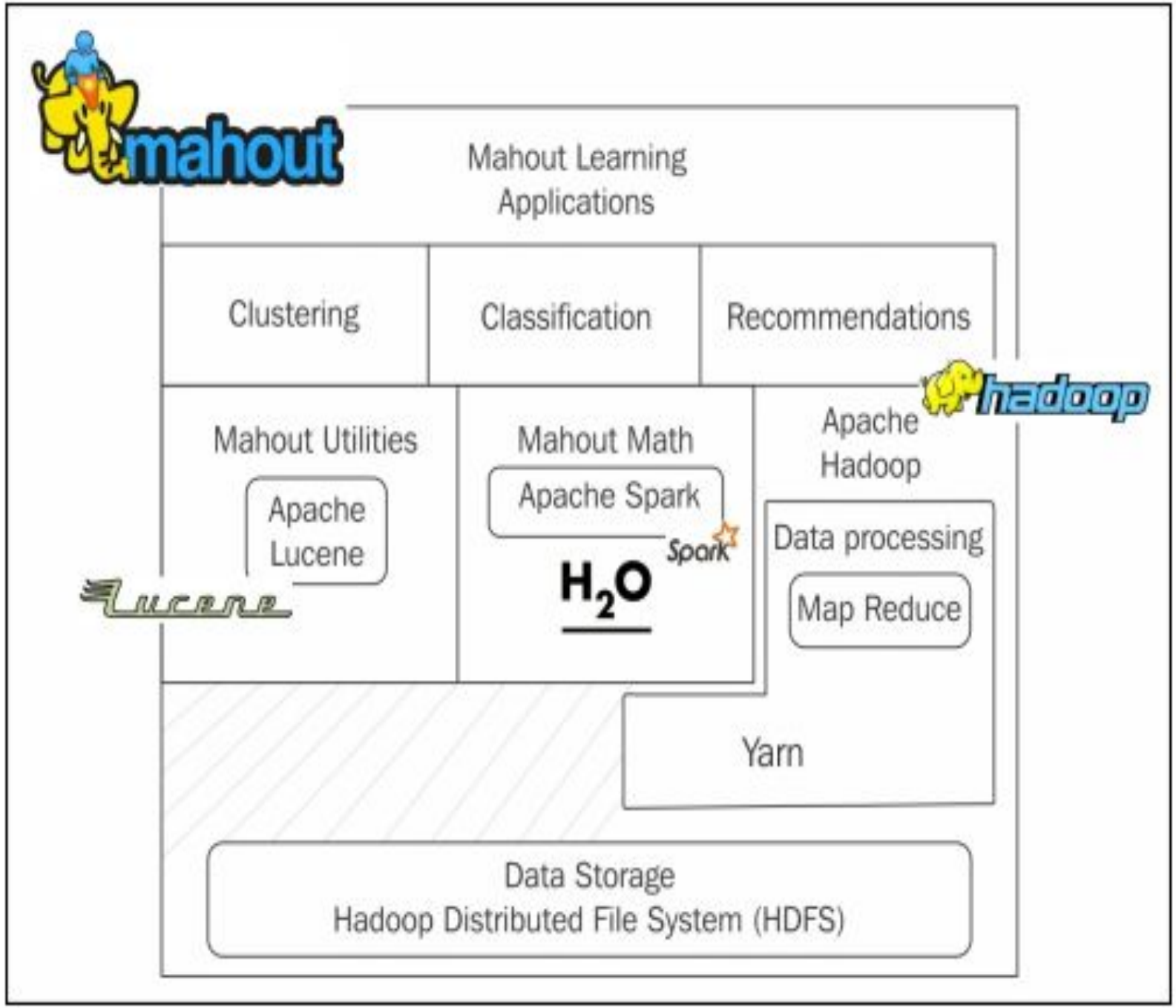
- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

# Mahout


- Hadoop is an open-source framework from Apache that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.
- Apache Mahout is an open source project that is primarily used for creating scalable machine learning algorithms. It implements popular machine learning techniques such as:
  - Recommendation
  - Classification
  - Clustering
- Apache Mahout started as a sub-project of Apache's Lucene in 2008. In 2010, Mahout became a top level project of Apache.

# Mahout

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783554997/5/ch05lvl1sec42/apache-mahout-with-hadoop



# Mahout- Mahout provides Java code that implements the algorithms for several techniques in the following three categories

- **Classification:**
  - Logistic regression
  - Naïve Bayes
  - Random forests
  - Hidden Markov models
- **Clustering:**
  - Canopy clustering
  - K-means clustering
  - Fuzzy k-means
  - Expectation maximization (EM)
- **Recommenders/collaborative filtering:**
  - Nondistributed recommenders
  - Distributed item-based collaborative filtering

# Mahout

Applications of Mahout

- Companies such as Adobe, Facebook, LinkedIn, Foursquare, Twitter, and Yahoo use Mahout internally.
- Foursquare helps you in finding out places, food, and entertainment available in a particular area. It uses the recommender engine of Mahout.
- Twitter uses Mahout for user interest modelling.
- Yahoo! uses Mahout for pattern mining.

# Syllabus

- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.

# NoSQL

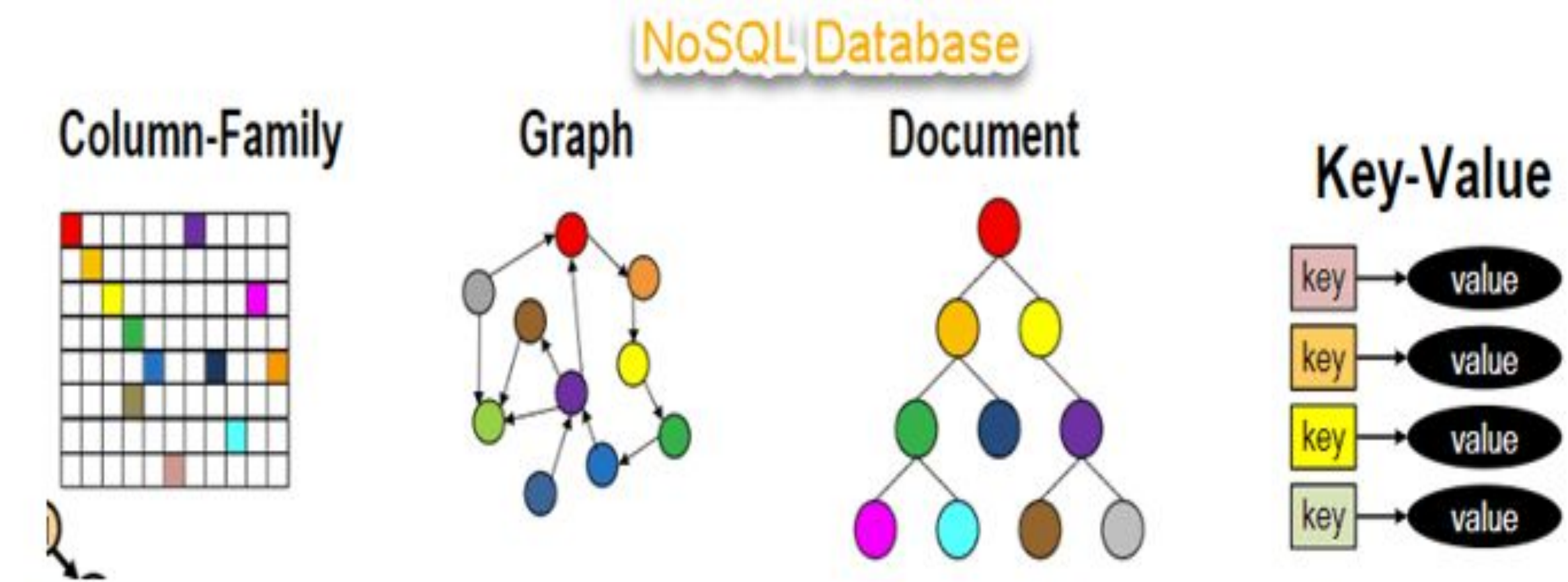
- NoSQL Database is a non-relational Data Management System, that does not require a fixed schema.
- It avoids joins, and is easy to scale.
- The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs.
- NoSQL is used for Big data and real-time web apps.
- For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day.
- NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.

# NoSQL Data store Types

There are four big NoSQL types:

1. key-value store.
2. document store.
3. column-oriented database.
4. graph database.

# NoSQL- Data store




https://medium.com/swlh/4-types-of-nosql-databases-d88ad21f7d3b

# NoSQL- Data store

## Key Value

Example:
Riak, Tokyo Cabinet, Redis server, Memcached, Scalaris

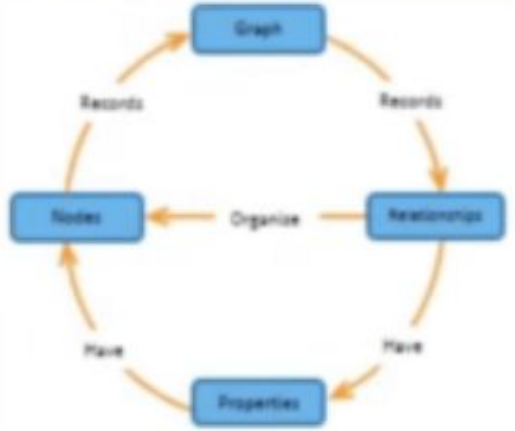
## Document-Based


Example:
MongoDB, CouchDB, OrientDB, RavenDB

## Column-Based

Example:
BigTable, Cassandra, Hbase, Hypertable

## Graph-Based




Example:
Neo4J, InfoGrid, Infinite Graph, Flock DB

# NoSQL- Key/value stores

- Key/value stores contain data (the value) that can be simply accessed by a given identifier (the key).
- The values can be complex.
- In a key/value store, there is no stored structure of how to use the data; the client that reads and writes to a key/value store needs to maintain and utilize the logic of how to meaningfully extract the useful elements from the key and the value.

- Here are some uses for key/value stores:
  - Using a customer's login ID as the key, the value contains the customer's preferences.
  - Using a web session ID as the key, the value contains everything that was captured during the session.

# NoSQL - Document stores

- Document stores are useful when the value of the key/value pair is a file and the file itself is self-describing (for example, JSON or XML).
- The underlying structure of the documents can be used to query and customize the display of the documents' content.
- Because the document is self-describing, the document store can provide additional functionality over a key/value store.
- For example, a document store may provide the ability to create indexes to speed the searching of the documents.
- Document stores may be useful for the following:
  - Content management of web pages
  - Web analytics of stored log data

# NoSQL -Column family

- Column family stores are useful for sparse datasets, records with thousands of columns but only a few columns have entries.
- The key/value concept still applies, but in this case a key is associated with a collection of columns.
- In this collection, related columns are grouped into column families.

- For example, columns for age, gender, income, and education may be grouped into a demographic family.
- Column family data stores are useful in the following instances:
  - To store and render blog entries, tags, and viewers' feedback
  - To store and update various web page metrics and counters

# NoSQL-Graph databases

- Graph databases are intended for use cases such as networks, where there are items (people or web page links) and relationships between these items.
- While it is possible to store graphs such as trees in a relational database, it often becomes cumbersome to navigate, scale, and add new relationships.
- Graph databases help to overcome these possible obstacles and can be optimized to quickly traverse a graph (move from one item in the network to another item in the network)
- Following are examples of graph database implementations:
  - Social networks such as Facebook and LinkedIn
  - Geospatial applications such as delivery and traffic systems to optimize the time to reach one or more destinations

# Pig Vs Hive

| Apache Pig | Hive |
| --- | --- |
| Apache Pig uses a language called **Pig Latin**. It was originally created at **Yahoo**. | Hive uses a language called **HiveQL**. It was originally created at **Facebook**. |
| Pig Latin is a data flow language. | HiveQL is a query processing language. |
| Pig Latin is a procedural language and it fits in pipeline paradigm. | HiveQL is a declarative language. |
| Apache Pig can handle structured, unstructured, and semi-structured data. | Hive is mostly for structured data. |

https://drive.google.com/file/d/1rwq28opUsOyNGFCl5cQVOsOyX7XEJj1p/view

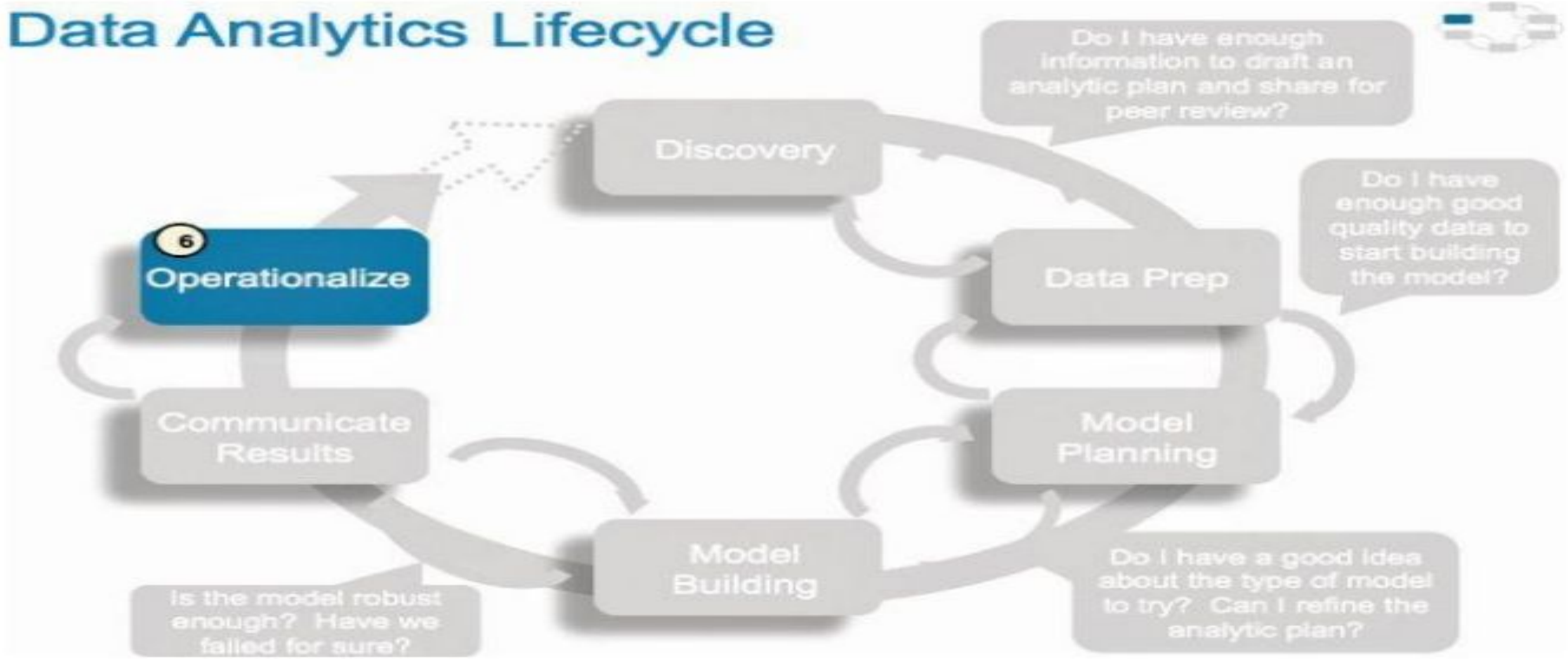
# Syllabus

- **Analytics for unstructured data-**
  - Use cases,
  - Map Reduce,
  - Apache Hadoop.
- The Hadoop Ecosystem-
  - Pig,
  - HIVE,
  - HBase,
  - Mahout,
  - NoSQL.
- An Analytics Project-
  - Communicating & operationalizing,
  - Creating final deliverables.

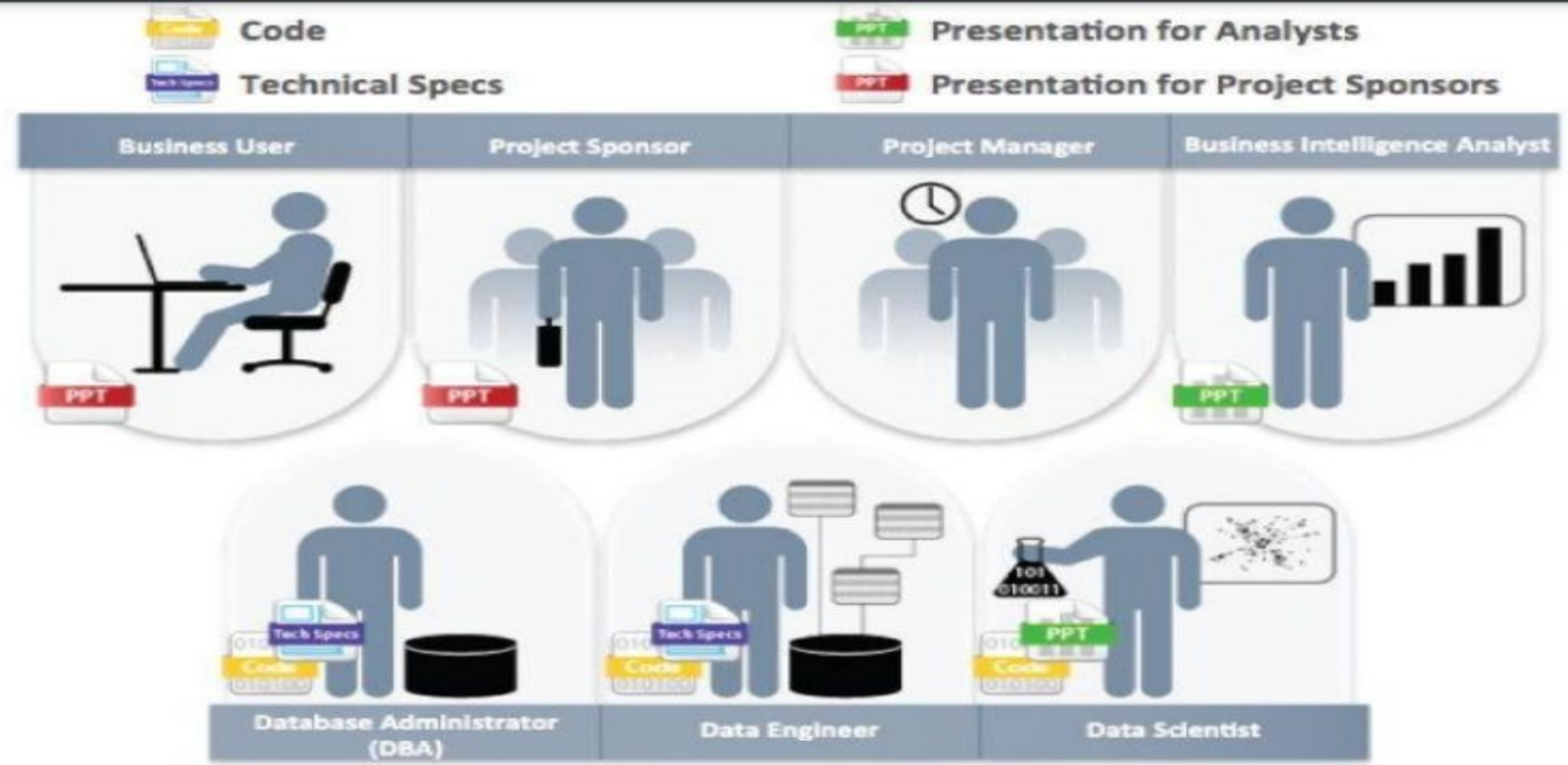
# Syllabus

- An Analytics Project-
  - Communicating & operationalizing
  - creating final deliverables.

# Operationalize



# Communicating and Operationalizing an Analytics Project- Key outputs from a successful analytic project



# key outputs for stakeholders of analytics project

- **Business User** typically tries to determine the benefits and implications of the findings to the business.
- **Project Sponsor** typically asks questions related to the business impact of the project, the risks and return on investment (ROI), and how the project can be evangelized within the organization and beyond.
- **Project Manager** needs to determine if the project was completed on time and within budget.
- **Business Intelligence Analyst** needs to know if the reports and dashboards he manages will be impacted and need to change.
- **Data Engineer and Database Administrator (DBA)** typically need to share the code from the analytical project and create technical documents that describe how to implement the code.
- **Data Scientists** need to share the code and explain the model to their peers,managers, and other stakeholders.

# four main deliverables

- **Presentation for Project Sponsors** contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- **Presentation for Analysts**, which describes changes to business processes and reports. Data scientists reading this presentation are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms) and will be interested in the details.
- **Code for technical people**, such as engineers and others managing the production environment
- **Technical specifications** for implementing the code

# Syllabus

- An Analytics Project-
  - Communicating & operationalizing
  - Creating final deliverables.

# Creating the Final Deliverables

- Synopsis of YoyoDyne Bank case study example
- Analytics plan for YoyoDyne Bank case study
- Developing Core Material for Multiple Audiences
- Project Goals
- Main Findings
- Approach
- Model Description
- Key Points Supported with Data
- Model Details
- Recommendations
- Additional Tips on the Final Presentation
- Providing Technical Specifications and Code

https://bhavanakhivsara.files.wordpress.com/2018/06/data-science-and-big-data-analy-nieizv_book.pdf
Point 12.2

# References


- **Book:** Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data : EMC Education Services.
- https://medium.com/swlh/4-types-of-nosql-databases-d88ad21f7d3b