# Introduction to Big Data

## ECAP456

Dr. Rajni Bhalla
Associate Professor

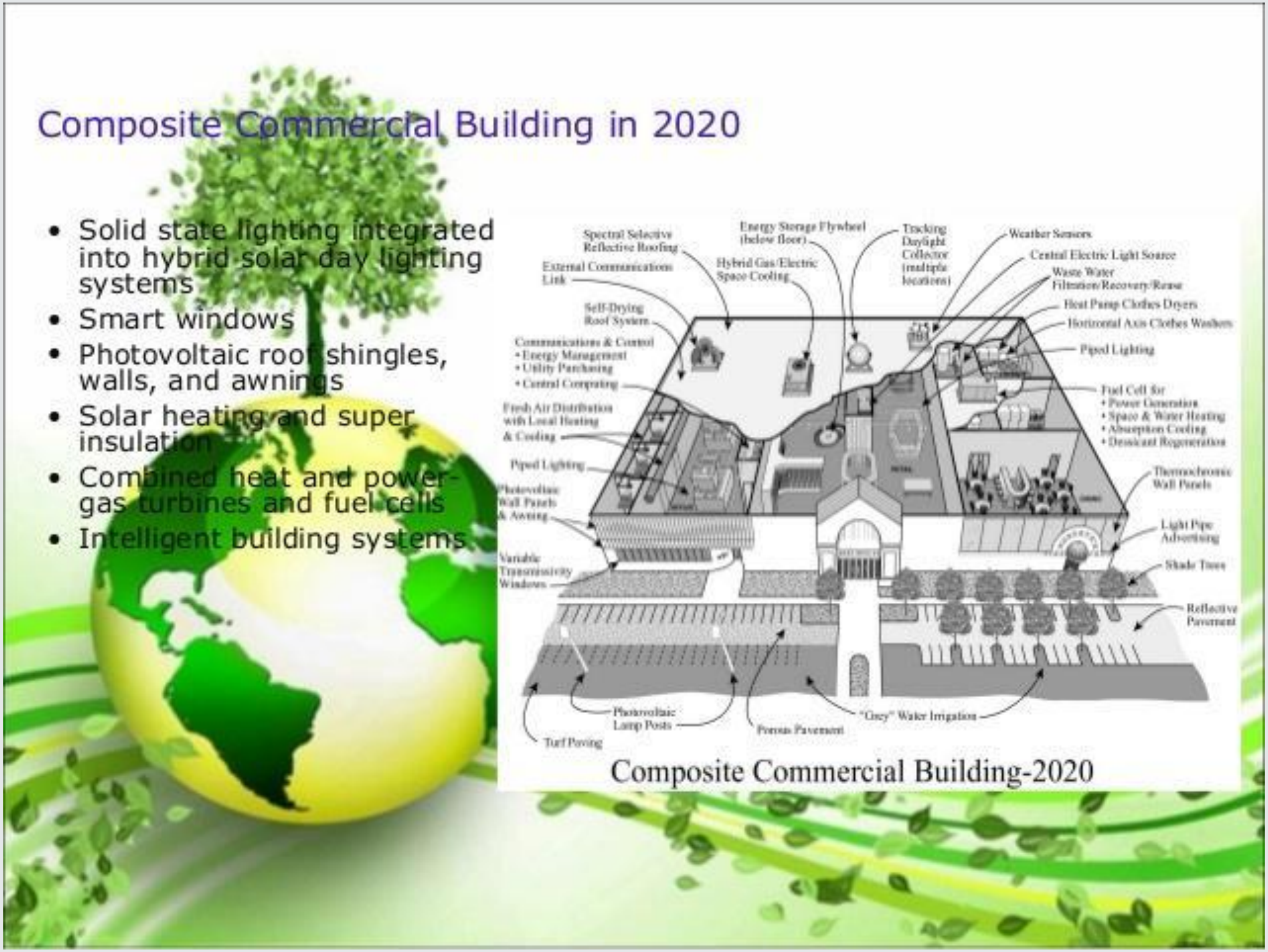
# Learning Outcomes

After this lecture, you will be able to

- learn Open-Source Software Related To Hadoop.

# Introduction

Energy efficient & Green Model



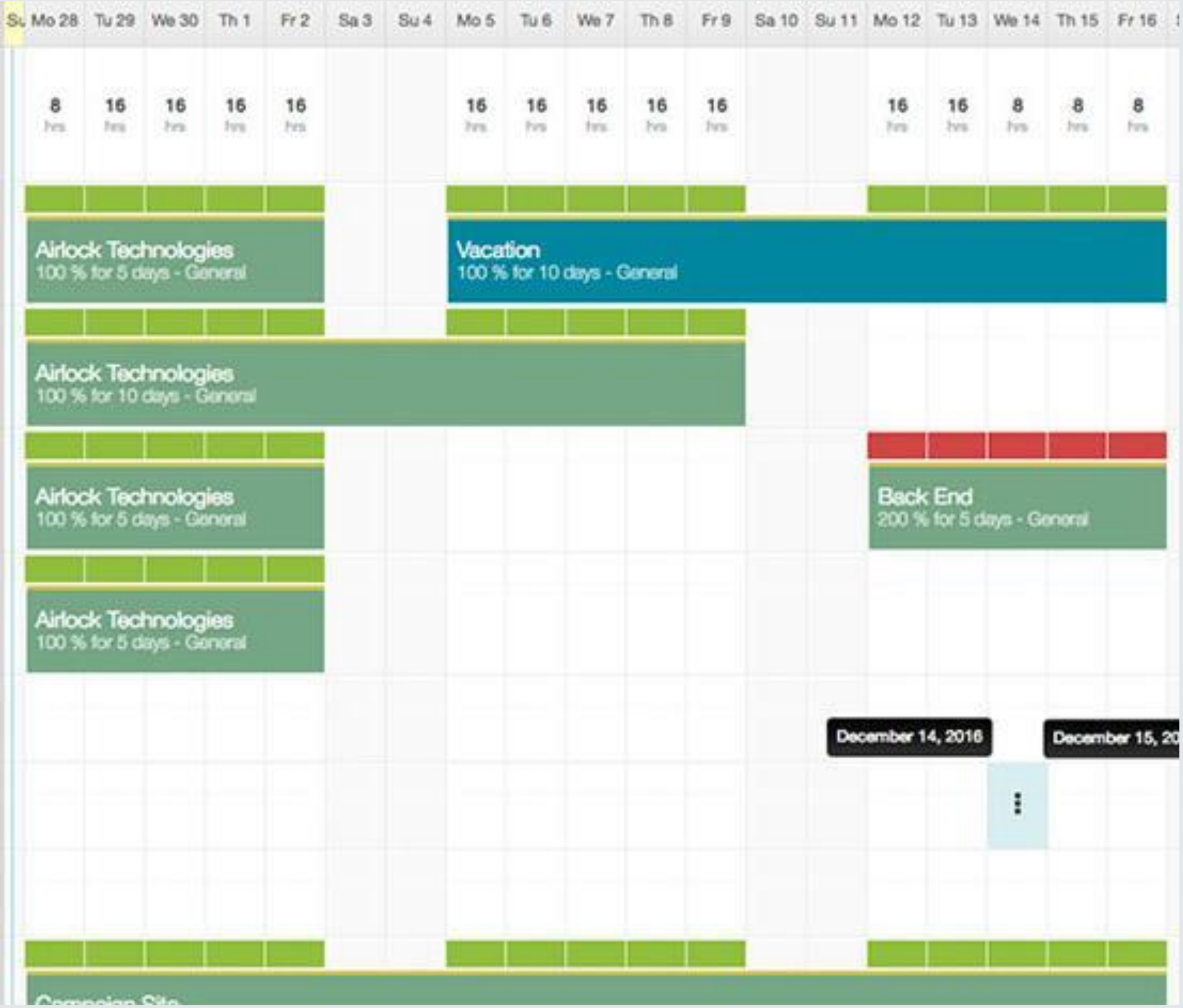
## Composite Commercial Building in 2020

- Solid state lighting integrated into hybrid solar day lighting systems
- Smart windows
- Photovoltaic roof shingles, walls, and awnings
- Solar heating and super insulation
- Combined heat and power-gas turbines and fuel cells
- Intelligent building systems

Spectral Selective Reflective Roofing
External Communications Link
Energy Storage Flywheel (below floor)
Hybrid Gas/Electric Space Cooling
Tracking Daylight Collector (multiple locations)
Weather Sensors
Central Electric Light Source
Waste Water Filtration/Recovery/Reuse
Heat Pump Clothes Dryers
Horizontal Axis Clothes Washers
Piped Lighting
Self-Drying Roof System
Communications & Control
• Energy Management
• Utility Purchasing
• Central Computing
Fresh Air Distribution with Local Heating & Cooling
Piped Lighting
Fuel Cell for
• Power Generation
• Space & Water Heating
• Absorption Cooling
• Desiccant Regeneration
Photovoltaic Wall Panels & Awning
Thermochromic Wall Panels
Light Pipe Advertising
Variable Transmissivity Windows
Shade Trees
Reflective Pavement
Photovoltaic Lamp Posts
Turf Paving
Porous Pavement
"Grey" Water Irrigation

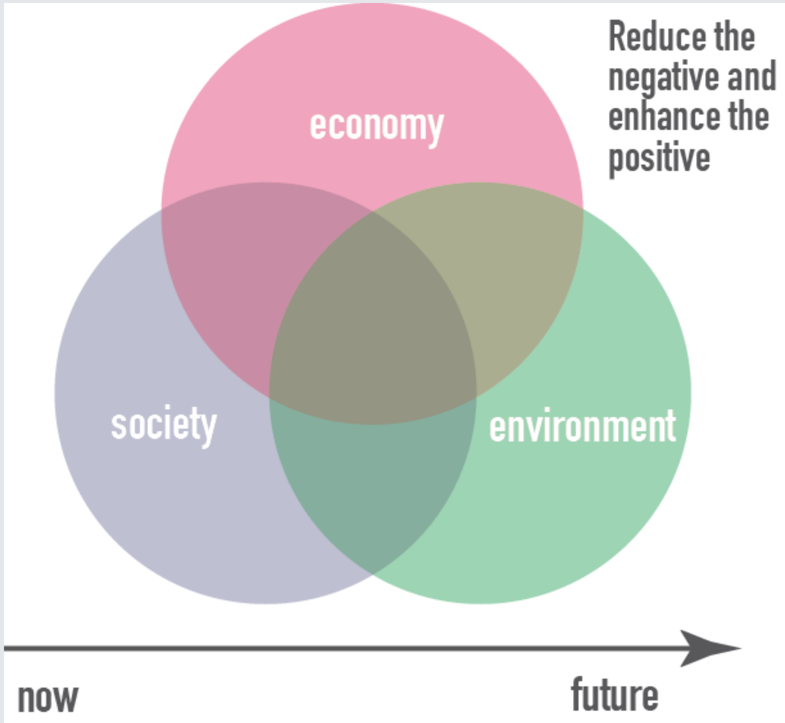
Composite Commercial Building-2020

# Introduction



Resource Scheduling

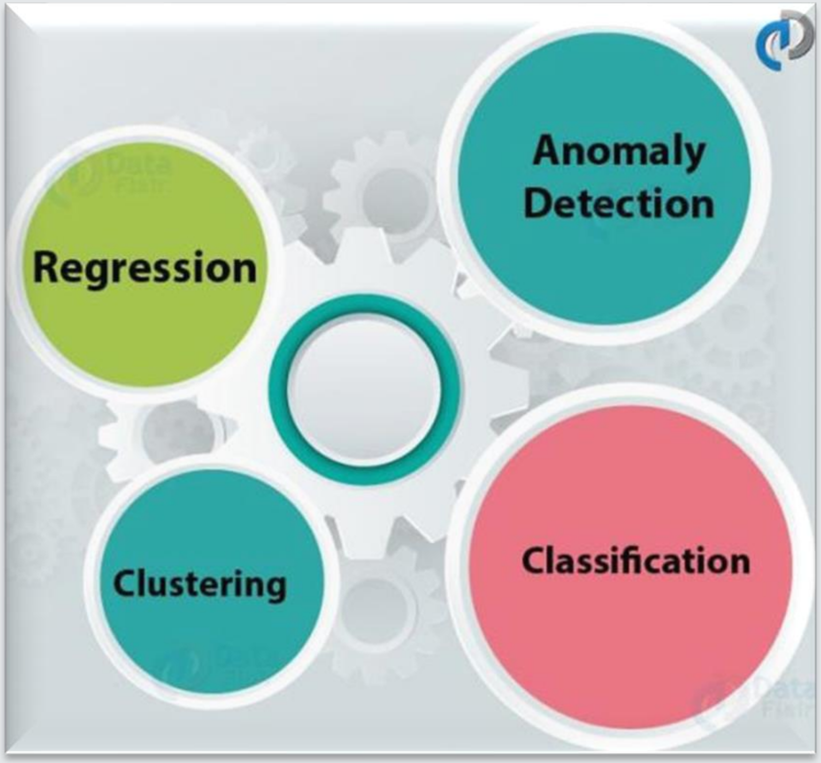
# Introduction



Sustainability Issues

Fault Tolerance and Reliability

# Introduction



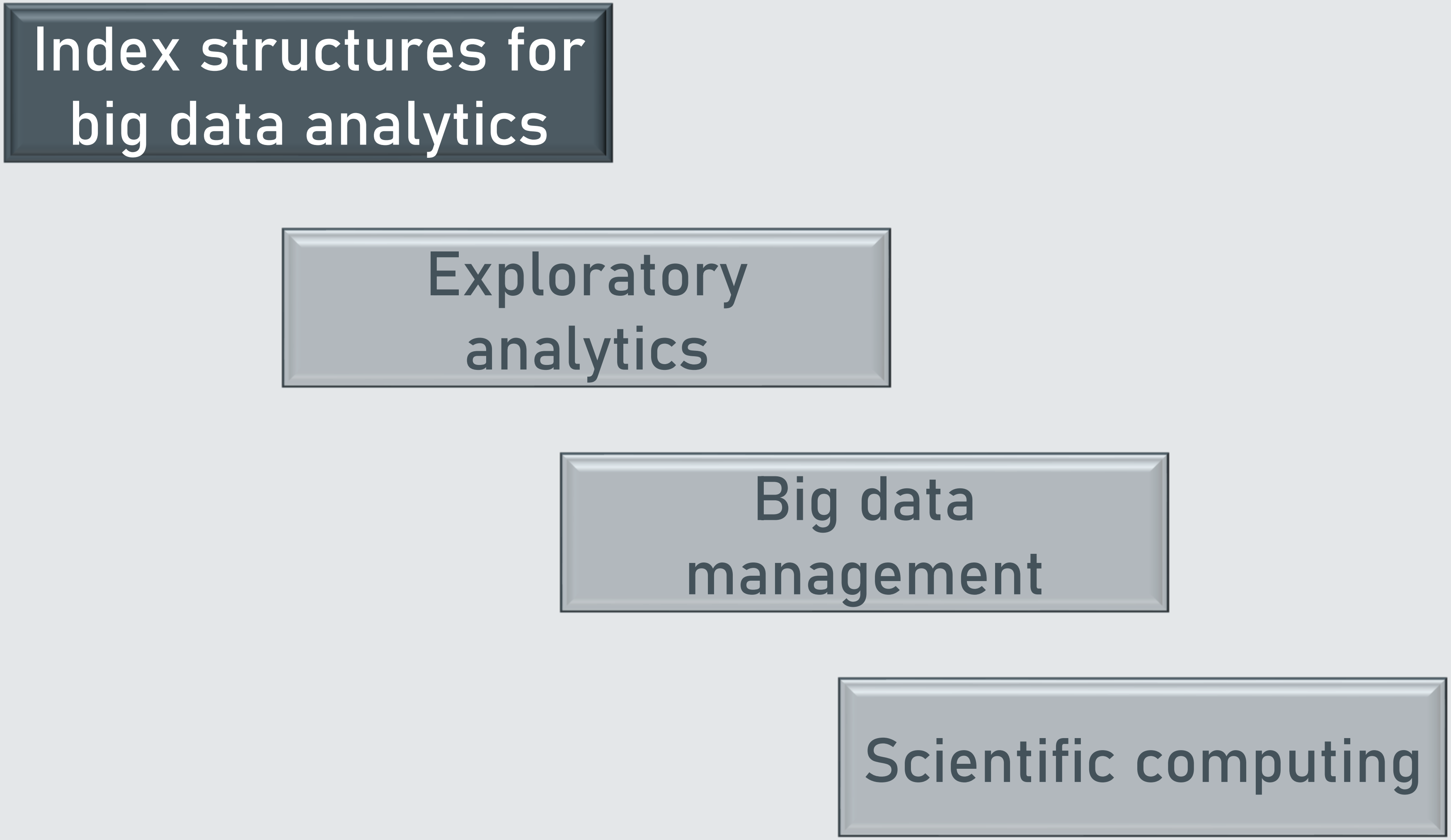
Machine Learning Techniques

Graph Analysis

# Introduction

Large Scale Recommender System

# Introduction



Index structures for big data analytics

Exploratory analytics

Big data management

Scientific computing

# Introduction

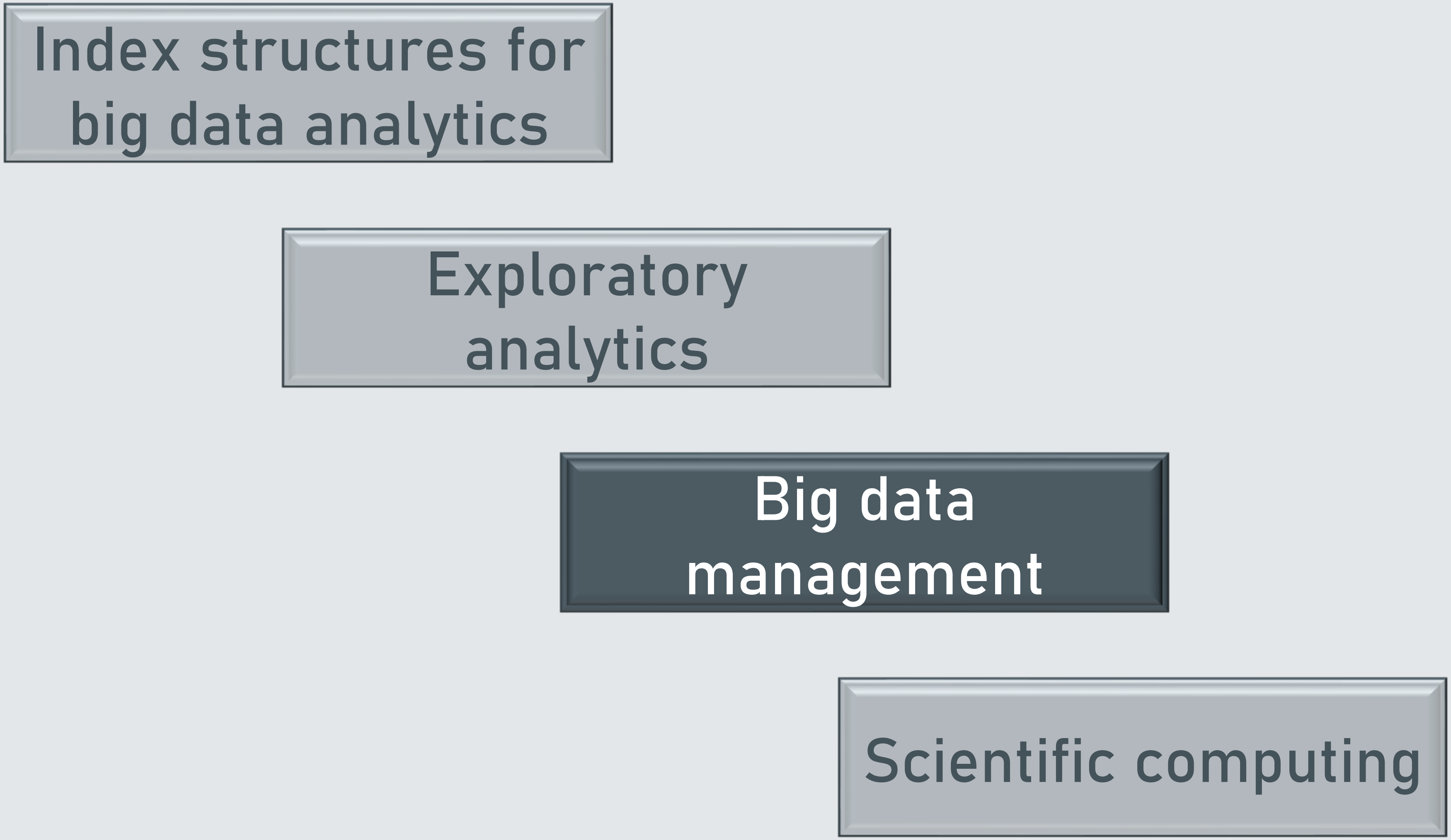
Index structures for big data analytics

Exploratory analytics

Big data management

Scientific computing

# Introduction



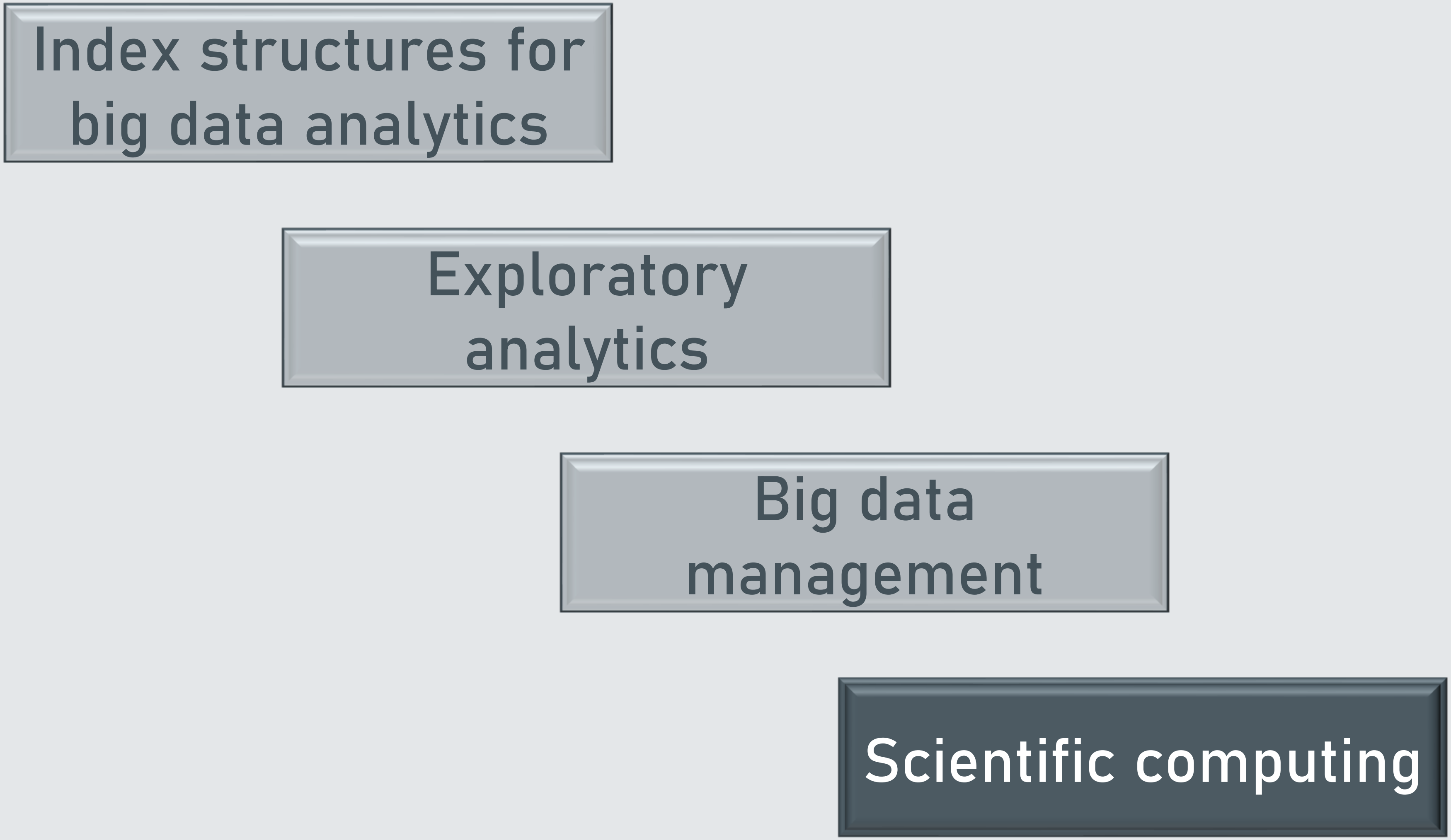
Index structures for big data analytics

Exploratory analytics

Big data management

Scientific computing

# Introduction



Index structures for big data analytics

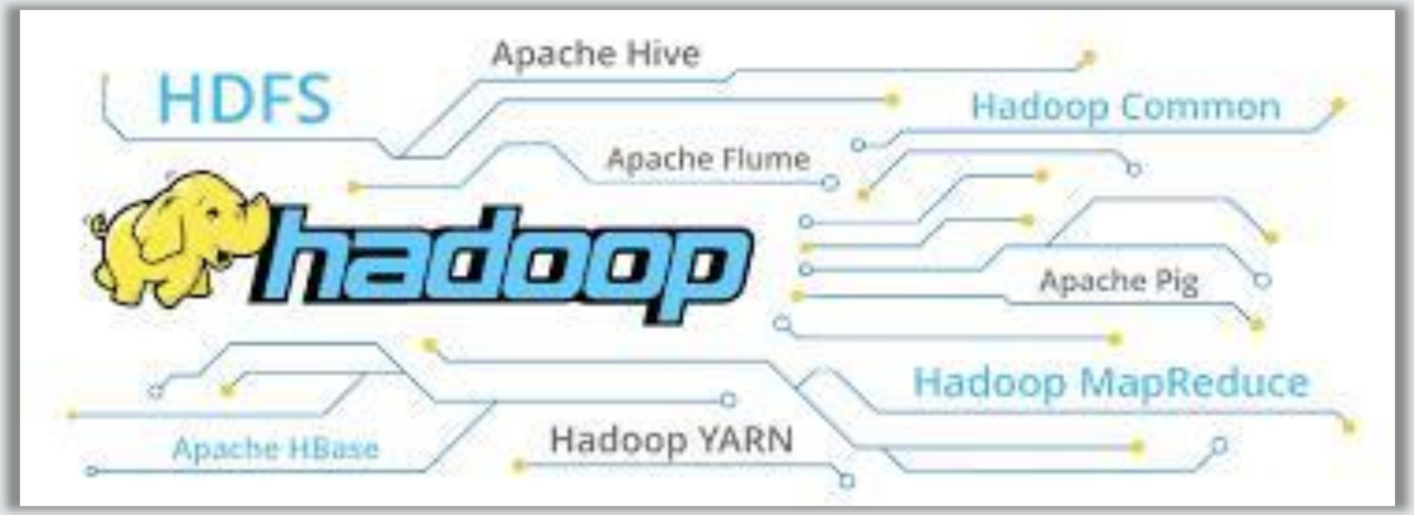
Exploratory analytics

Big data management

Scientific computing

# Hadoop

- Hadoop is an Open-Source Tool that available in public.
- It is a framework that provides too many services like Pig, Impala, Hive, HBase, etc.



# Top Hadoop Related Open-Source Tools

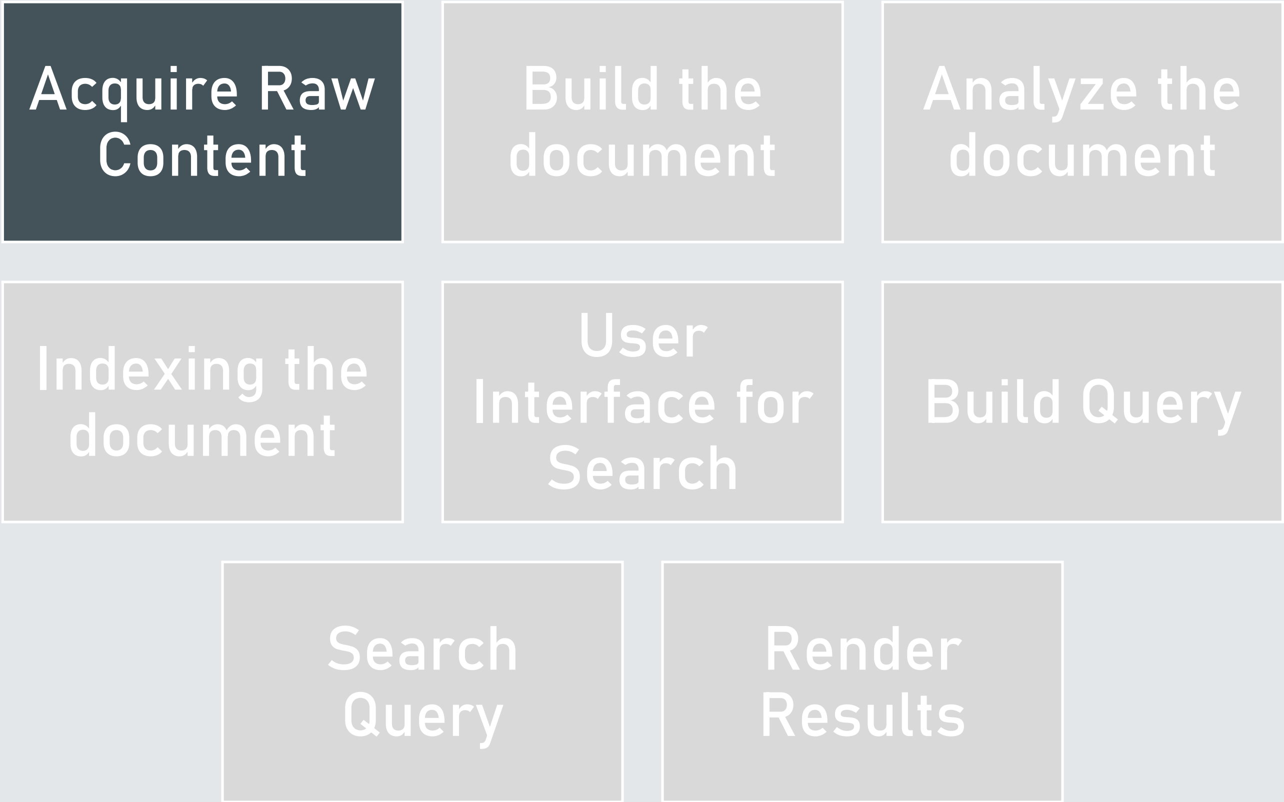
- Lucene [Java based text search engine]
- Eclipse [Popular IDE written in Java]
- HBase [Hadoop NoSQL Database]
- Hive [Query data engine]
- Jaql [Query language for JavaScript]
- Pig [Large datasets analyzing platform]
- Zookeeper [Centralized configuration service]
- Avro [Data serialization system]
- UIMA [unstructured analytic framework]
- Presto [Distributed SQL query solution]

# Lucene(Java Based Text Search Engine)

Lucene is an open source Java based search library. It is very popular and a fast search library. It is used in Java based applications to add document search capability to any kind of application in a very simple and efficient way.

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

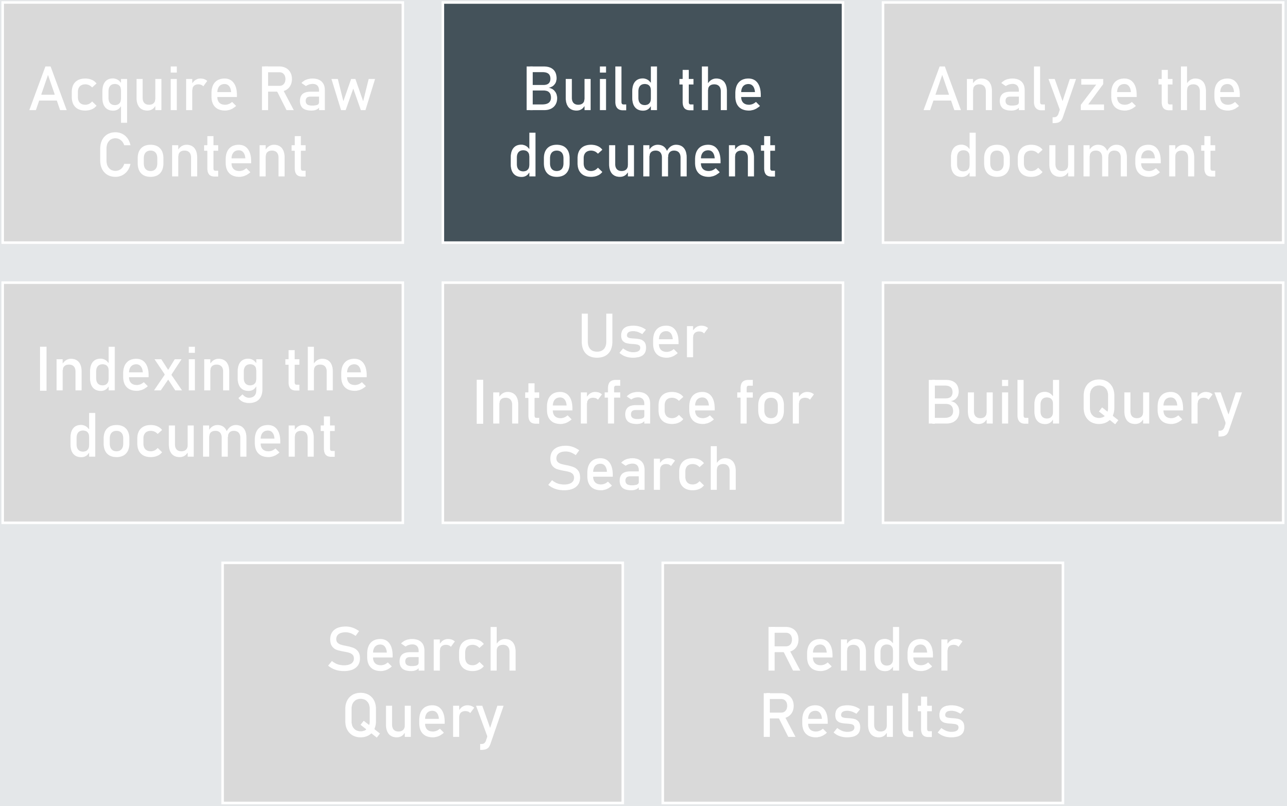
Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

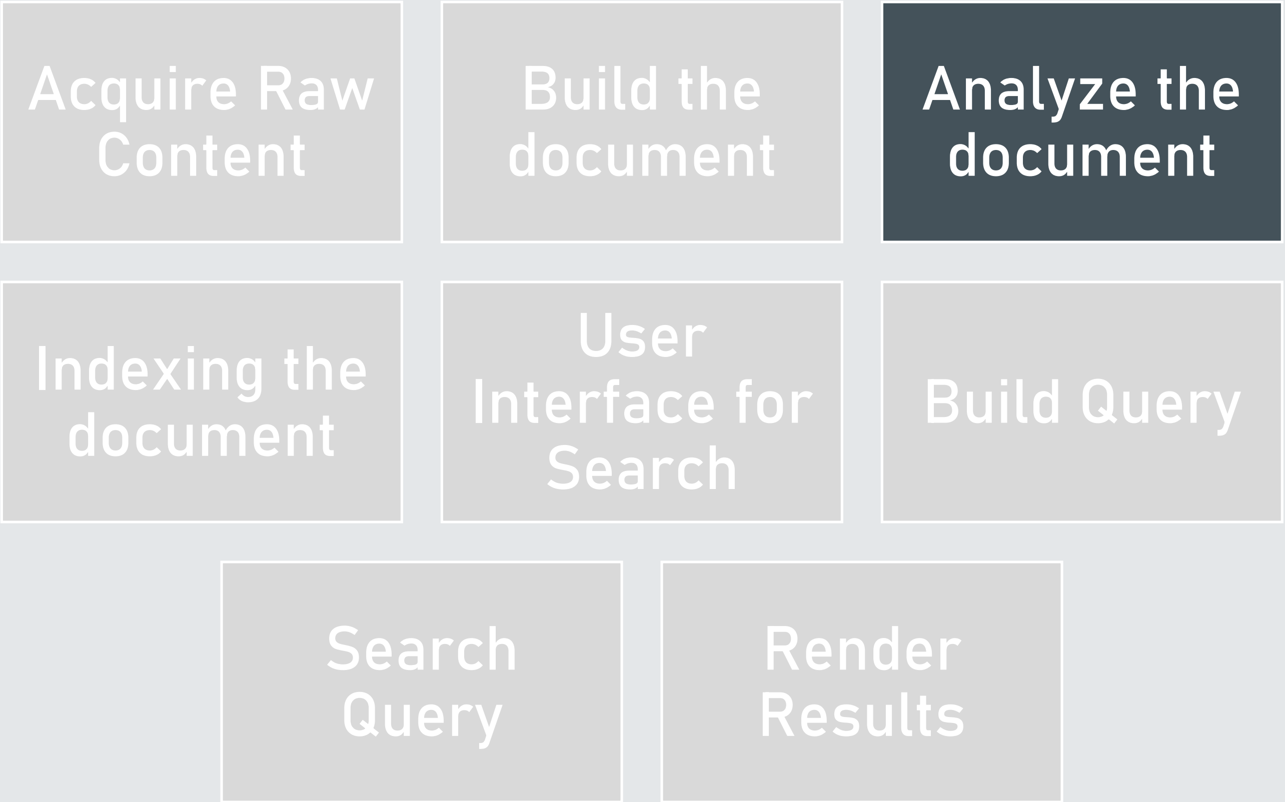
Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

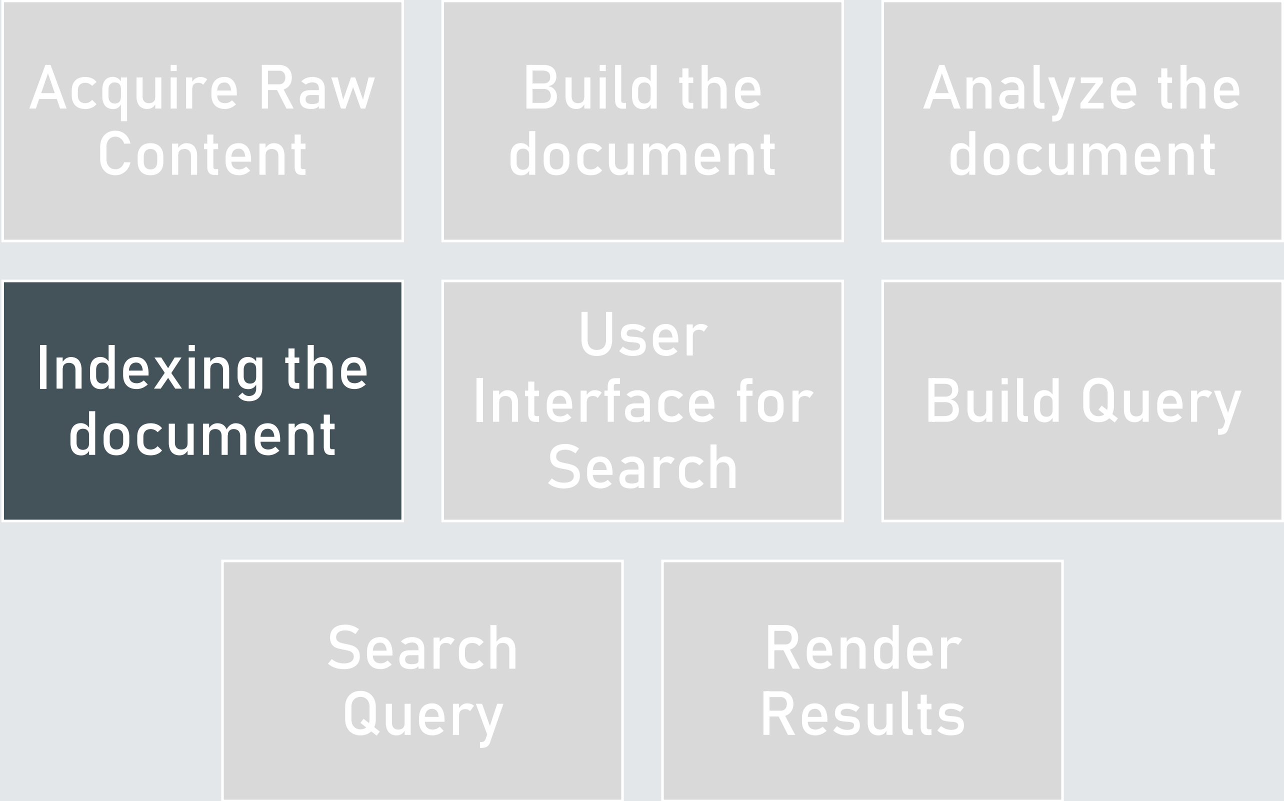
Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

**Indexing the document**

User Interface for Search

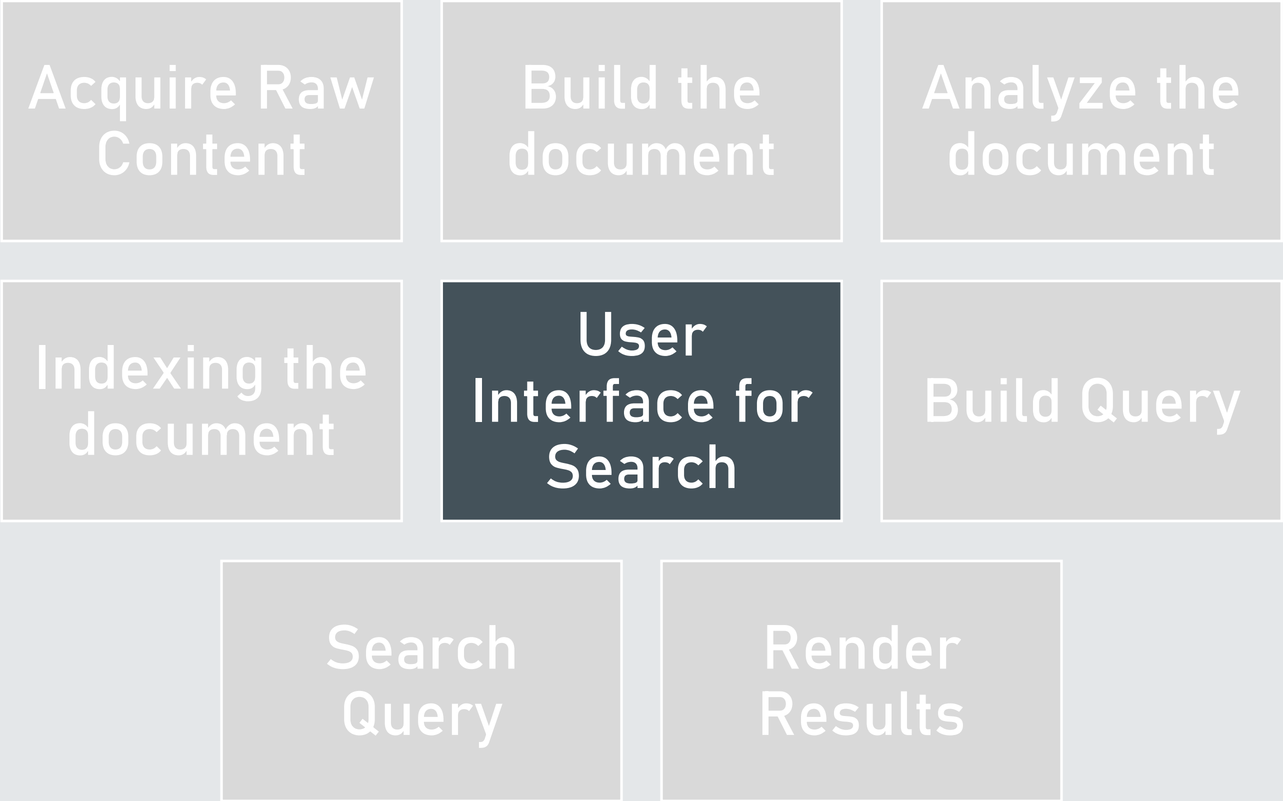
Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

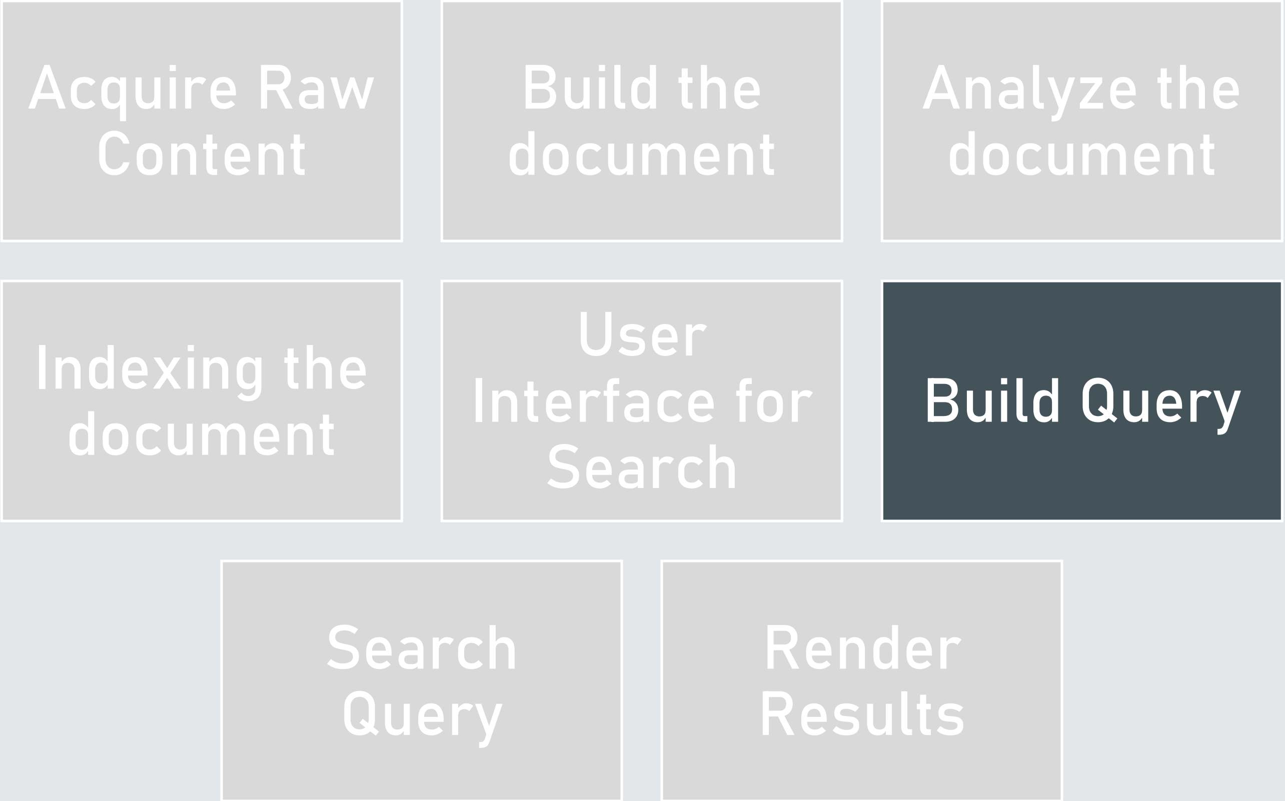
Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

Build Query

Search Query

Render Results

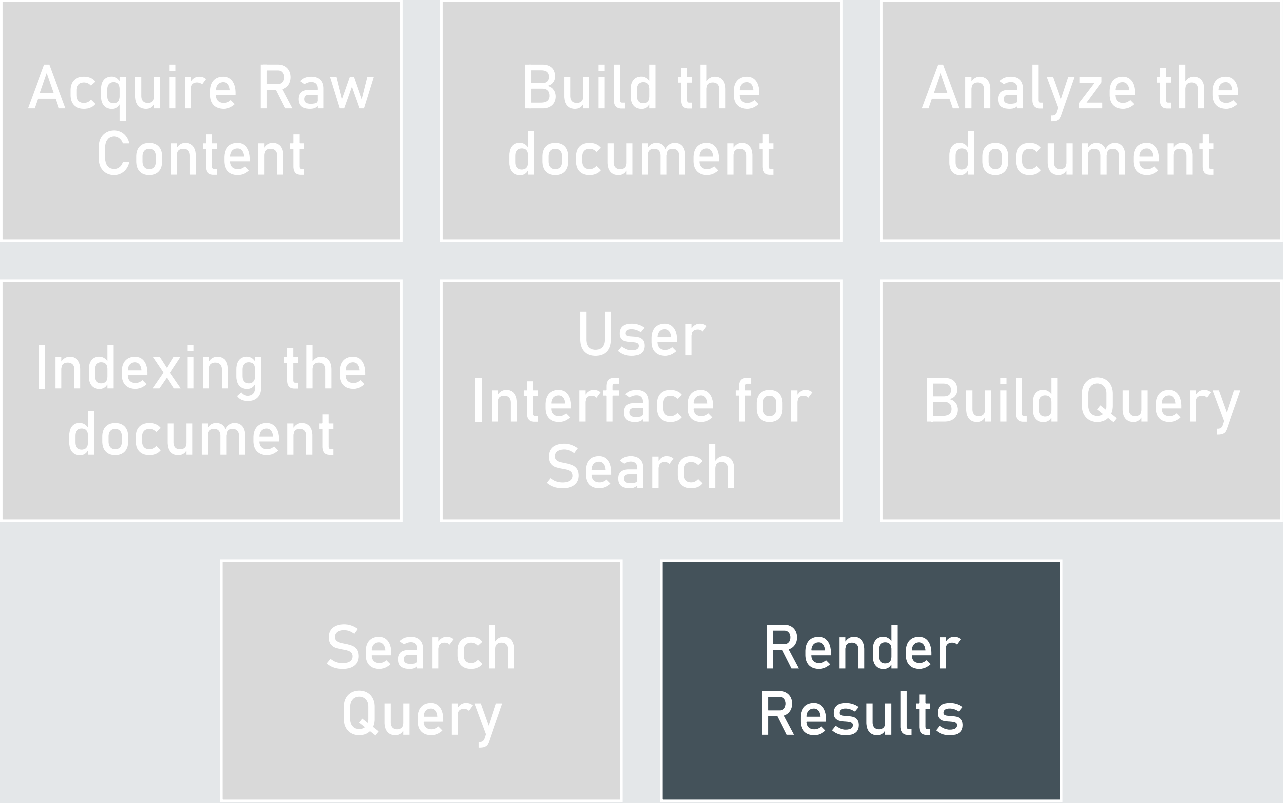
# Lucene(Java Based Text Search Engine)

## How Search Application works?

- Acquire Raw Content
- Build the document
- Analyze the document
- Indexing the document
- User Interface for Search
- Build Query
- **Search Query**
- Render Results.

# Lucene(Java Based Text Search Engine)

## How Search Application works?



Acquire Raw Content

Build the document

Analyze the document

Indexing the document

User Interface for Search

Build Query

Search Query

Render Results

# Lucene(Java Based Text Search Engine)

Apart from these basic operations, a search application can also provide administration user interface and help administrators of the application to control the level of search based on the user profiles. Analytics of search results is another important and advanced aspect of any search application.

# Eclipse [Popular IDE written in Java]

Eclipse is a Java IDE that is one of the 3 biggest and most popular IDE's in the world. It was written mostly in Java but it can also be used to develop applications in other programming languages apart from Java using plug-ins

Java Development using Eclipse IDE
© www.SoftwareTestingHelp.com

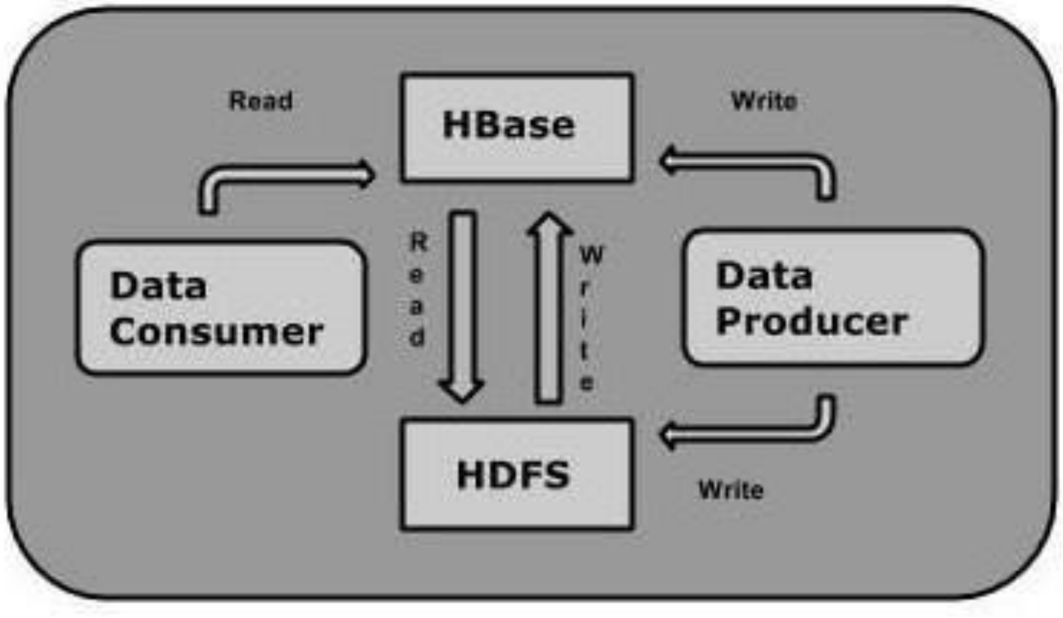
# Eclipse [Popular IDE written in Java]

Java Development using Eclipse IDE
© www.SoftwareTestingHelp.com

- PDE (Plugin Development Environment)
- Eclipse flaunts powerful tools for the various processes in application development .
- Eclipse can also be used to create various mathematical documents with LaTeX
- Eclipse can be used on platforms like Linux, macOS, Solaris and Windows.

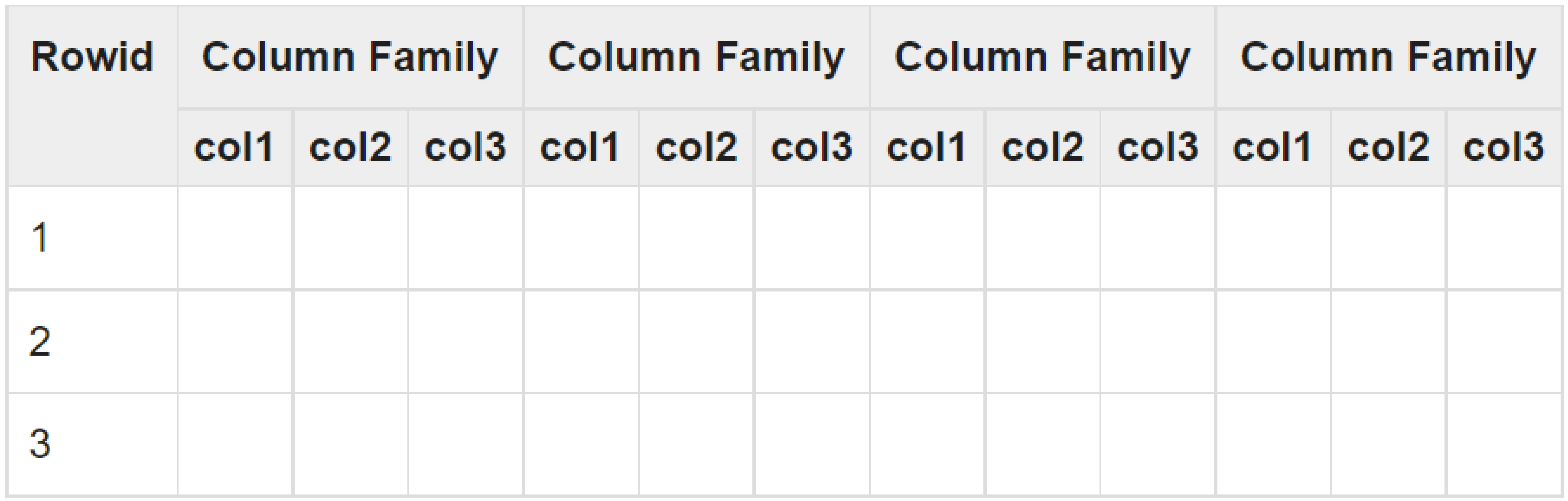
# HBase [Hadoop NoSQL Database]

HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data.



# HBase

## Storage Mechanism in HBase



| Rowid | Column Family | | | Column Family | | | Column Family | | | Column Family | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | col1 | col2 | col3 | col1 | col2 | col3 | col1 | col2 | col3 | col1 | col2 | col3 |
| 1 | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | |

# Hive [Query data engine]

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

# Hive [Query data engine]

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

# Hive [Query data engine]

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

# Hive [Query data engine]

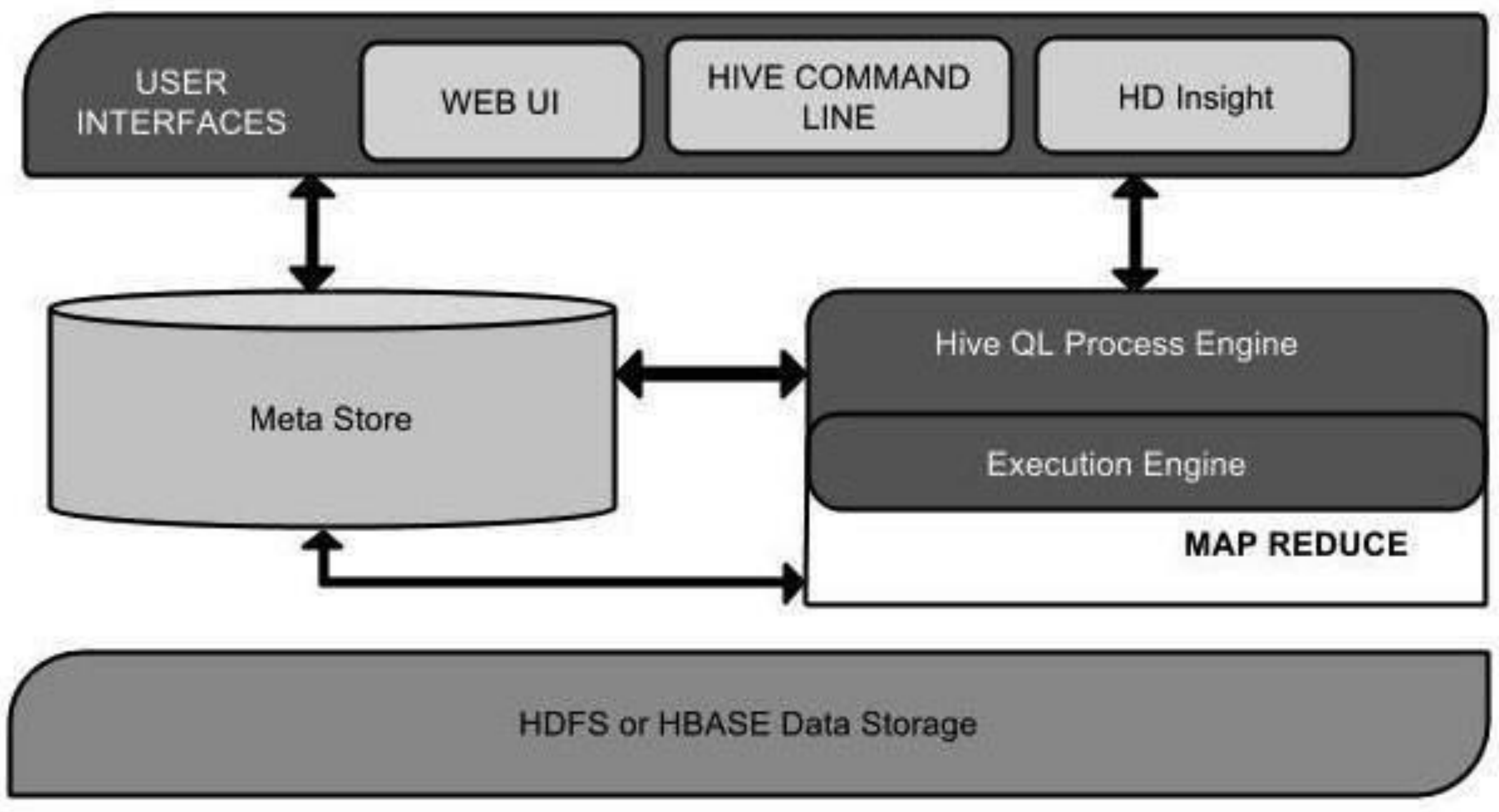
Hive is not

- A relational database
- A design for OnLine Transaction Processing (OLTP)
- A language for real-time queries and row-level updates

# Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.

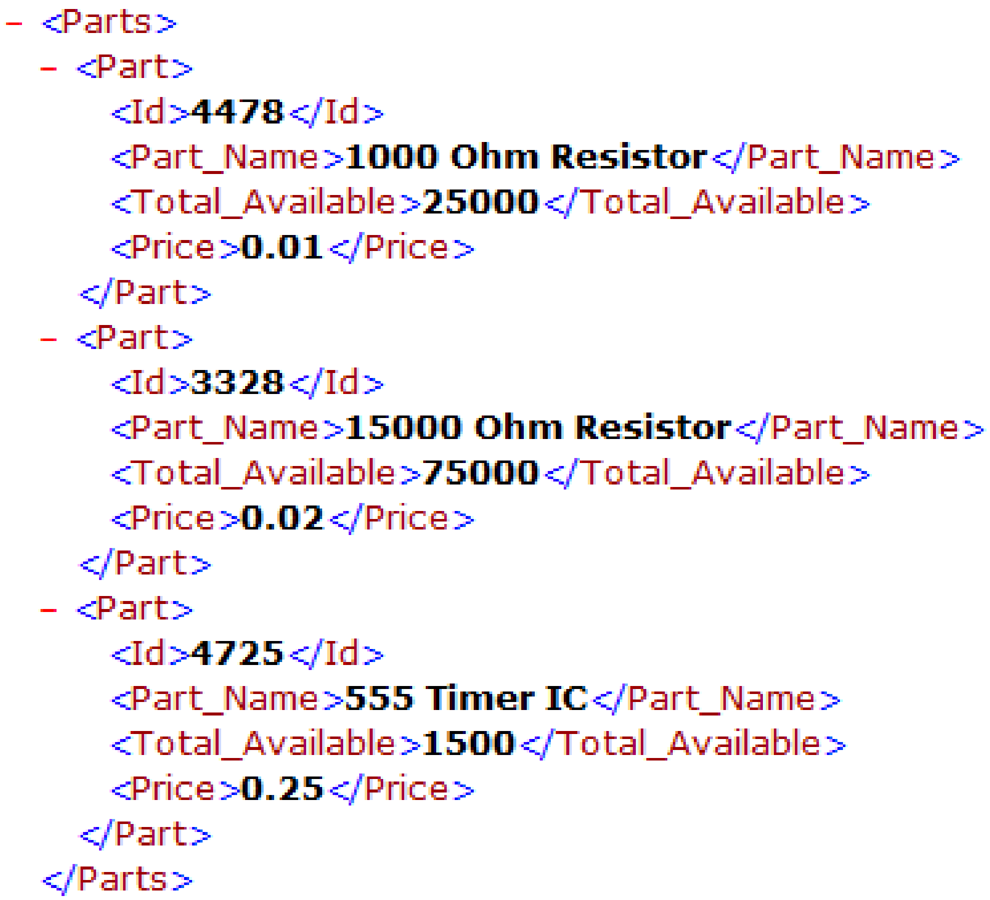
# Architecture of Hive



# Jaql [Query language for JavaScript]

JAQL is a query language for the JavaScript Object Notation (JSON) data interchange format. As its name implies, a primary use of JAQL is to handle data stored as JSON documents, but JAQL can work on various types of data.

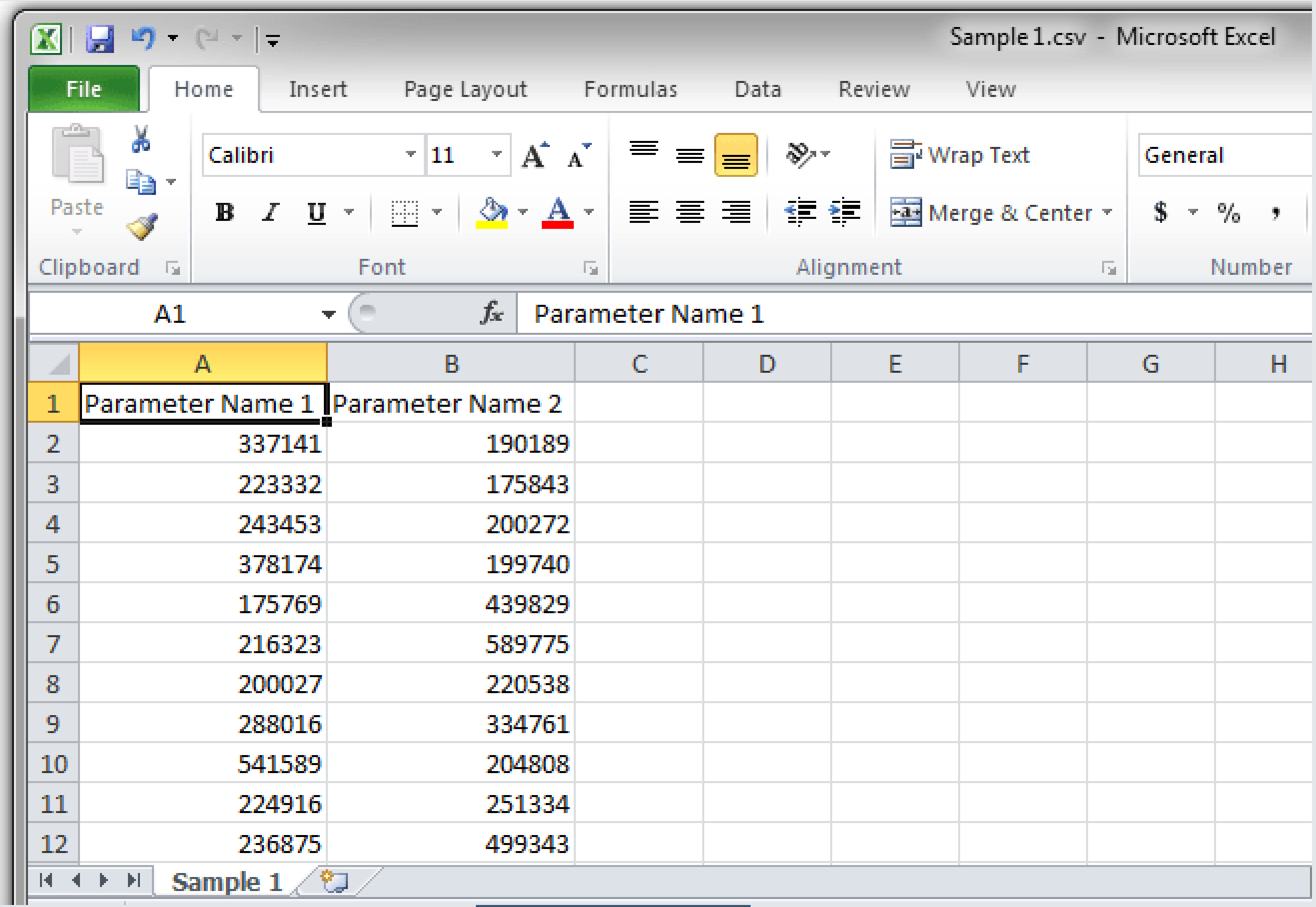
# Jaql [Query language for JavaScript]



```
- <Parts>
  - <Part>
      <Id>4478</Id>
      <Part_Name>1000 Ohm Resistor</Part_Name>
      <Total_Available>25000</Total_Available>
      <Price>0.01</Price>
    </Part>
  - <Part>
      <Id>3328</Id>
      <Part_Name>15000 Ohm Resistor</Part_Name>
      <Total_Available>75000</Total_Available>
      <Price>0.02</Price>
    </Part>
  - <Part>
      <Id>4725</Id>
      <Part_Name>555 Timer IC</Part_Name>
      <Total_Available>1500</Total_Available>
      <Price>0.25</Price>
    </Part>
  </Parts>
```

XML

# Jaql [Query language for JavaScript]



A1 | Parameter Name 1

| | A | B |
|---|---|---|
| 1 | Parameter Name 1 | Parameter Name 2 |
| 2 | 337141 | 190189 |
| 3 | 223332 | 175843 |
| 4 | 243453 | 200272 |
| 5 | 378174 | 199740 |
| 6 | 175769 | 439829 |
| 7 | 216323 | 589775 |
| 8 | 200027 | 220538 |
| 9 | 288016 | 334761 |
| 10 | 541589 | 204808 |
| 11 | 224916 | 251334 |
| 12 | 236875 | 499343 |

CSV

# Jaql [Query language for JavaScript]

Flat Files

Structured SQL data

# Jaql [Query language for JavaScript]

- JSON has found wide use in Web and mobile applications, including large-scale big data and enterprise data warehouse applications.
- JAQL can run in local mode on individual systems and in cluster mode, in the latter case supporting Hadoop applications. It automatically generates MapReduce jobs and parallel queries on Hadoop systems.

# Jaql [Query language for JavaScript]

IBM Research Labs

# Jaql [Query language for JavaScript]

Google Code

# Jaql [Query language for JavaScript]

Google Code

Apache 2.0 version

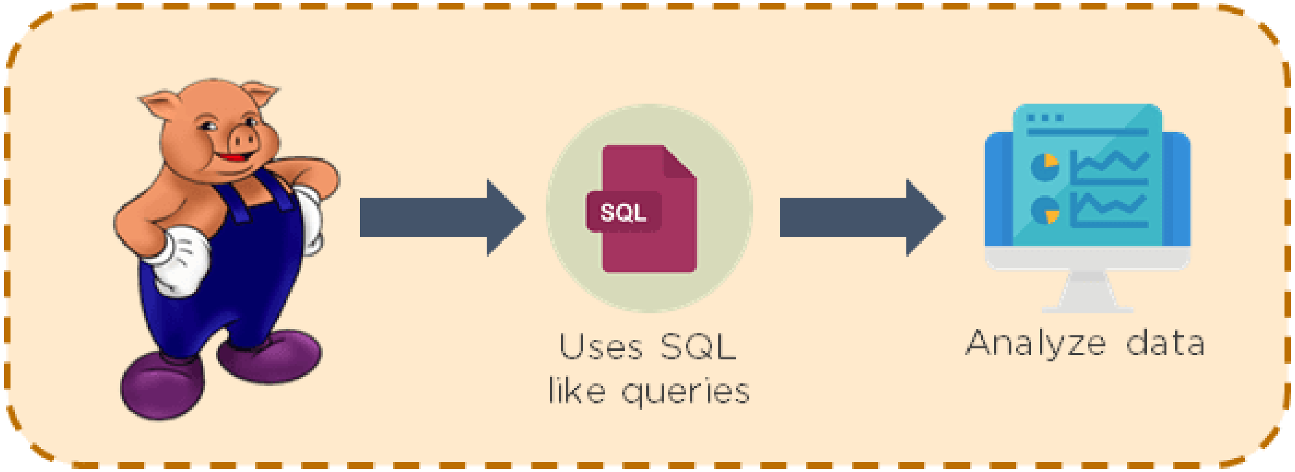
# Jaql [Query language for JavaScript]

Google Code

Apache 2.0 version

InfoSphere

# Pig [Large datasets analyzing platform]



Uses SQL like queries

Analyze data

Pig operates on various types of data like structured, semi-structured and unstructured data

# Pig [Large datasets analyzing platform]

## Why Do We Need Apache Pig?

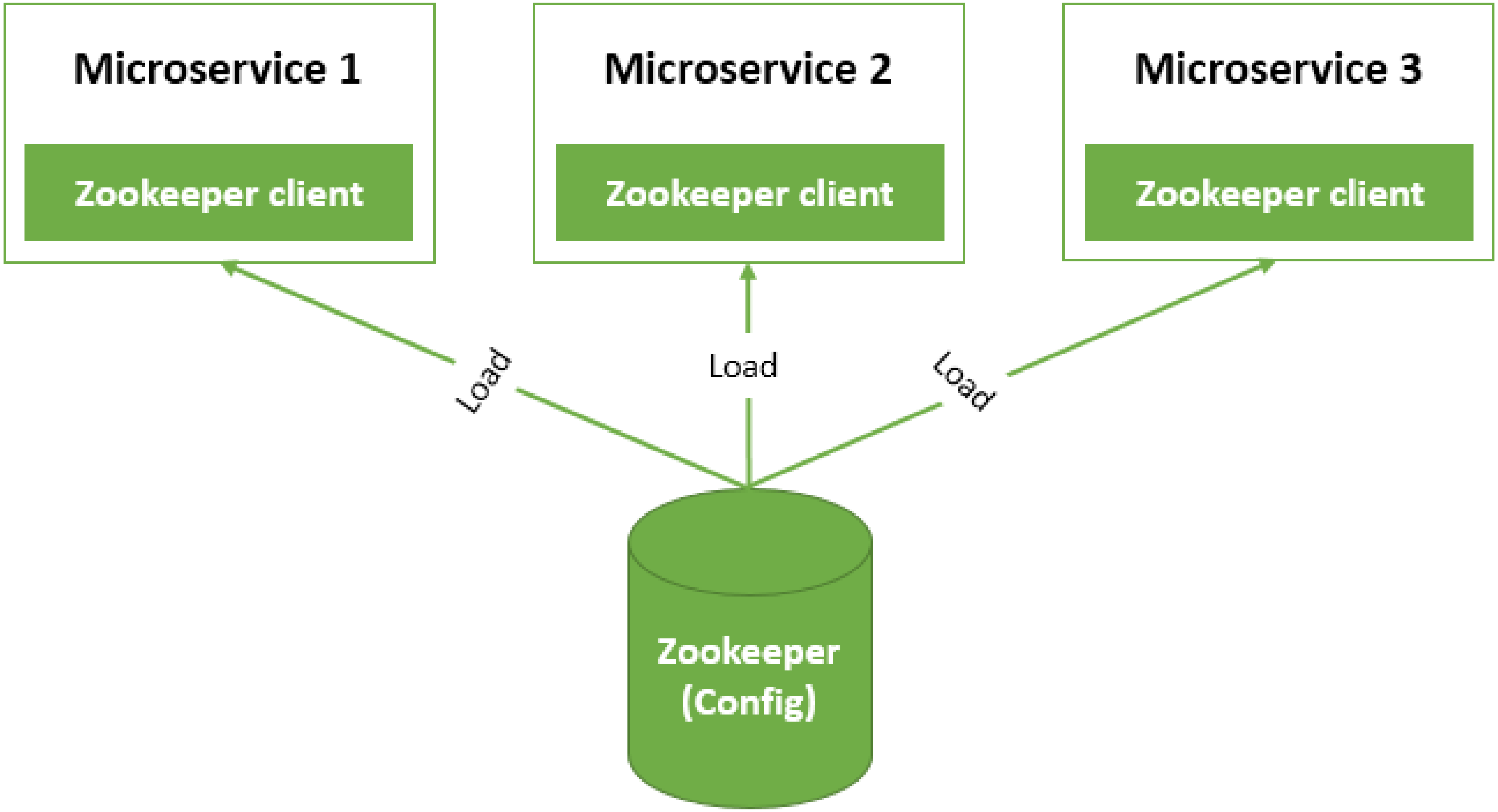
- Pig Latin
- multi-query approach,
- SQL-like language
- Built-in operators

# Pig [Large datasets analyzing platform]

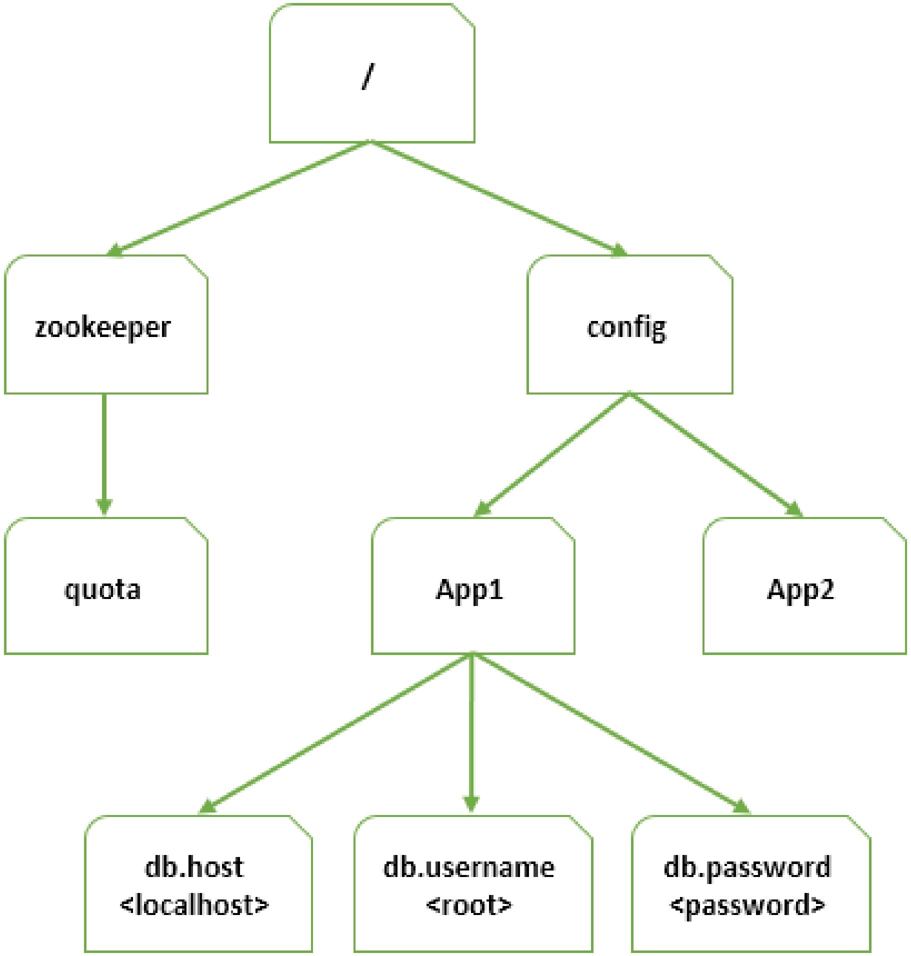
## Features of Pig

- Rich set of operators
- Ease of programming
- Optimization opportunities
- Extensibility
- User-defined Functions
- Handles all kinds of data

# Zookeeper [Centralized configuration service]



# Zookeeper [Centralized configuration service]



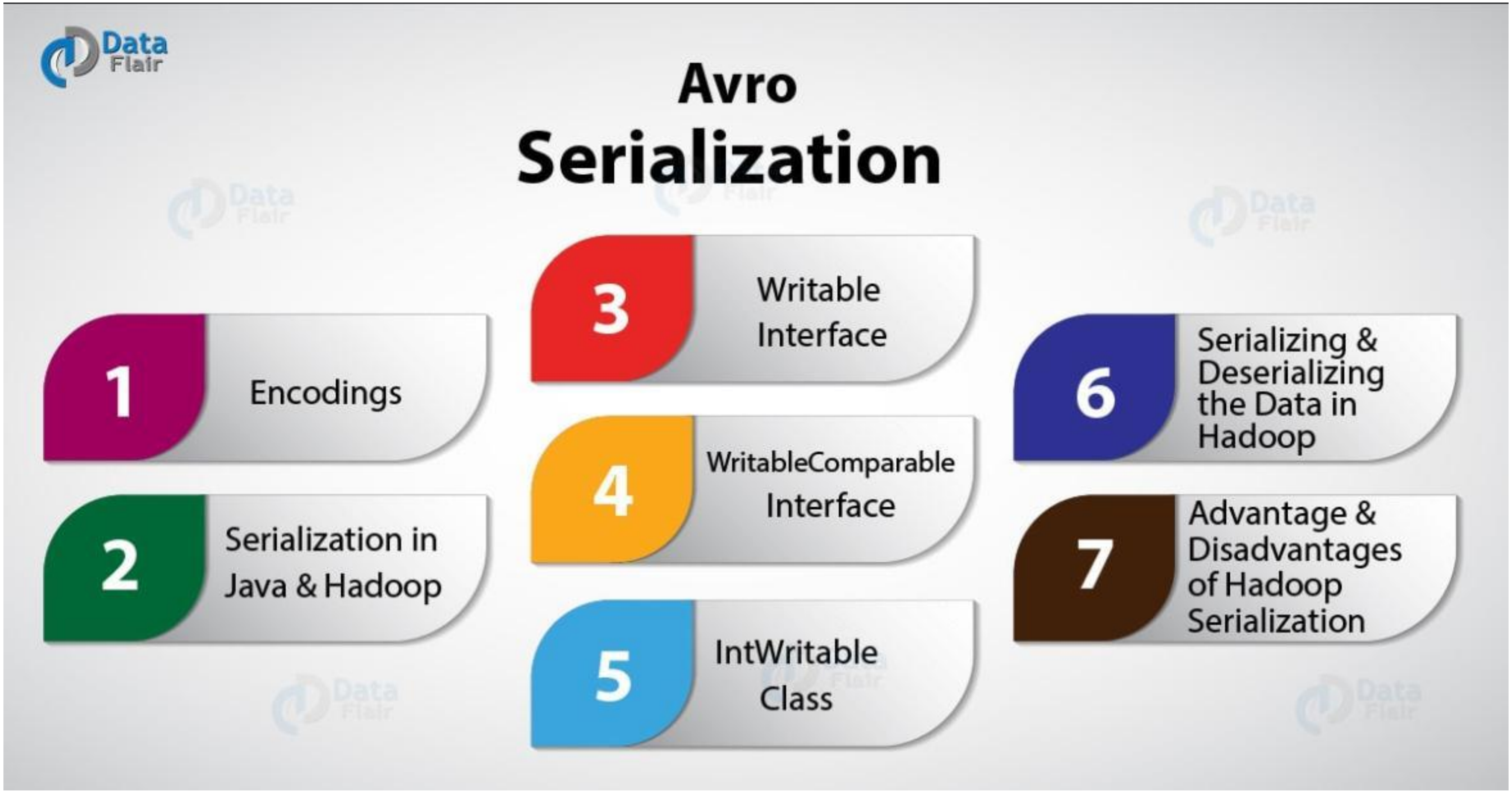
# Zookeeper [Centralized configuration service]

- This is a simplified version of how we are going to setup Zookeeper.
- Zookeeper stores data in a tree of ZNodes similar to Linux file system structure, a ZNode may contain another ZNodes or may have a value.
- App1 and App2 are sharing data from / and config znodes.
- However db.host, db.username and db.password are specific to App1.

# Avro [Data serialization system]

Doug Cutting

# Avro [Data serialization system]


Data Flair
Avro
Serialization
1 Encodings
2 Serialization in Java & Hadoop
3 Writable Interface
4 WritableComparable Interface
5 IntWritable Class
6 Serializing & Deserializing the Data in Hadoop
7 Advantage & Disadvantages of Hadoop Serialization

# Avro [Data serialization system]

Apache Avro is a language-neutral data serialization system. It was developed by Doug Cutting, the father of Hadoop. Since Hadoop writable classes lack language portability, Avro becomes quite helpful, as it deals with data formats that can be processed by multiple languages. Avro is a preferred tool to serialize data in Hadoop.

# Avro [Data serialization system]

Avro has a schema-based system. A language-independent schema is associated with its read and write operations. Avro serializes the data which has a built-in schema. Avro serializes the data into a compact binary format, which can be deserialized by any application.

# Avro [Data serialization system]

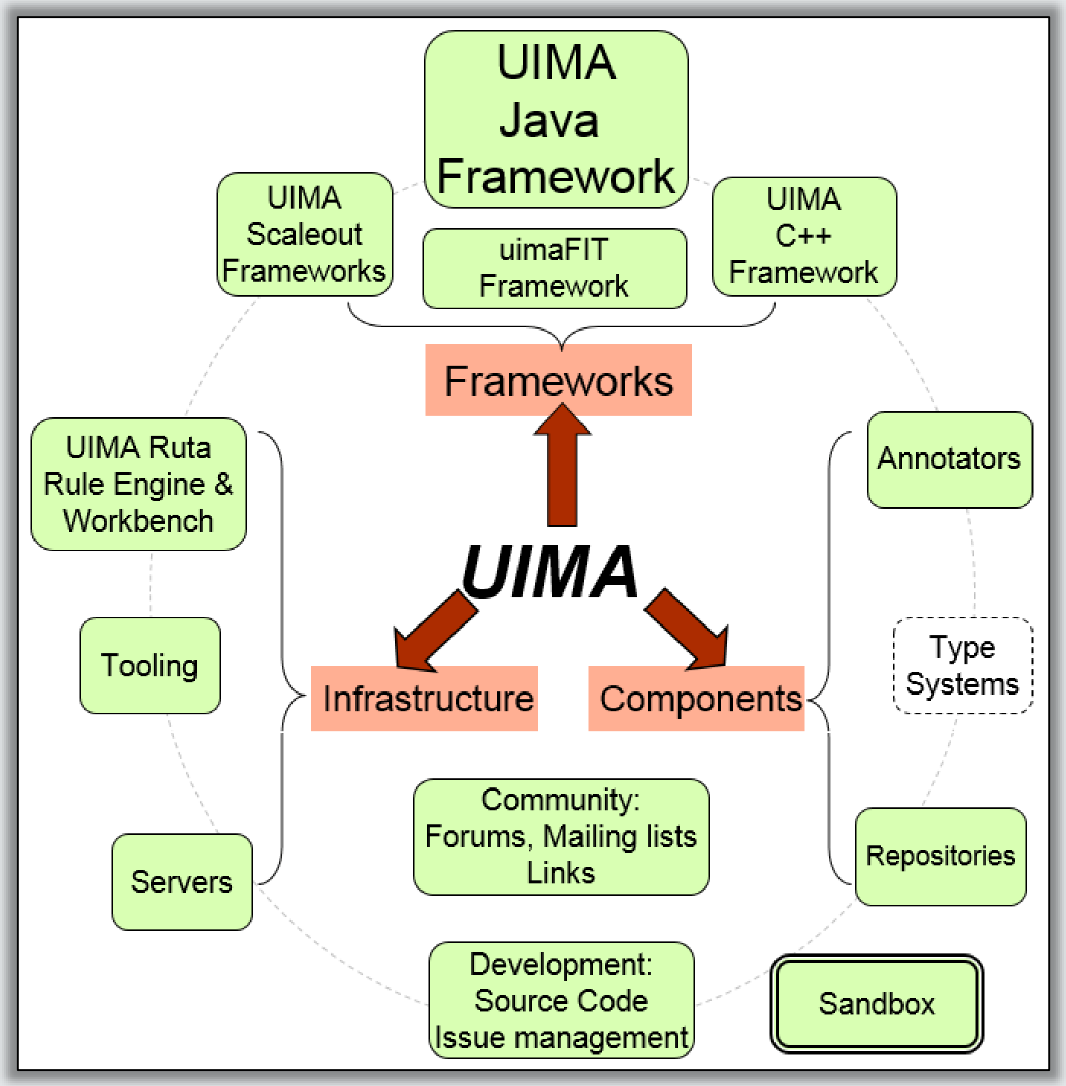
Avro uses JSON format to declare the data structures. Presently, it supports languages such as Java, C, C++, C#, Python, and Ruby.

# Avro [Data serialization system]

## Features of AVRO

- language-neutral
- processed by many languages
- compressible and splittable.
- rich data structures
- Avro schemas defined in JSON
- self-describing file named *Avro Data File*
- Remote Procedure Calls (RPCs).

# UIMA [unstructured analytic framework]



# UIMA [unstructured analytic framework]

Unstructured Information Management applications are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user. An example UIM application might ingest plain text and identify entities, such as persons, places, organizations; or relations, such as works-for or located-at.

# Presto [Distributed SQL query solution]

Presto (or PrestoDB) is an open source, distributed SQL query engine, designed from the ground up for fast analytic queries against data of any size.

# Presto [Distributed SQL query solution]

# Presto [Distributed SQL query solution]

That’s all for now…