02- Introduction to Big Data

Dr Shafaq Khan
School of Computer Science
University of Windsor

Submission Deadline

Project groups: Please add project lead details by highlighting it in the share document. Also, add an extra line below group title, specifying the project title.

Project proposal submission deadline: Sec 2: Jan 30; Sec 3: Jan 31; Sec 4: Feb 1

Agenda

- Evolution of Data
- Data Classification
- Introduction to Big data
- Big Data Analytics (BDA)
- The Hadoop Ecosystem
- Apache Spark
- The Architecture of NoSQL Databases
- **Workshop on MongoDB**

Introductory Questions

?? Why is the amount of data increasing tremendously?

?? How is data classified?

?? What is Big Data?

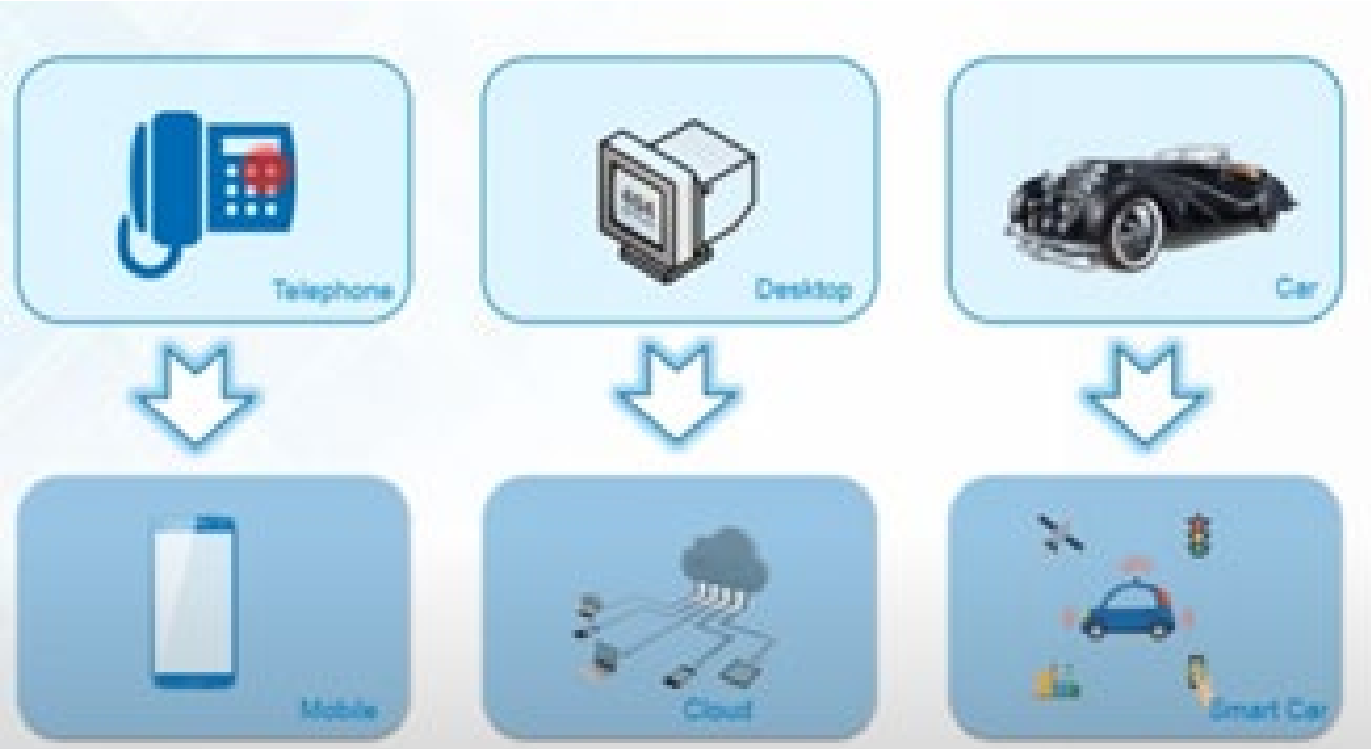
?? Why RDBMS is not suitable for Big Data?

?? Why Big Data Analysis is helpful?

?? How do we store and process big data?

Evolution of Data

Evolution of technology



As of November 2022, 59.5% of all website traffic comes from people using mobile devices.

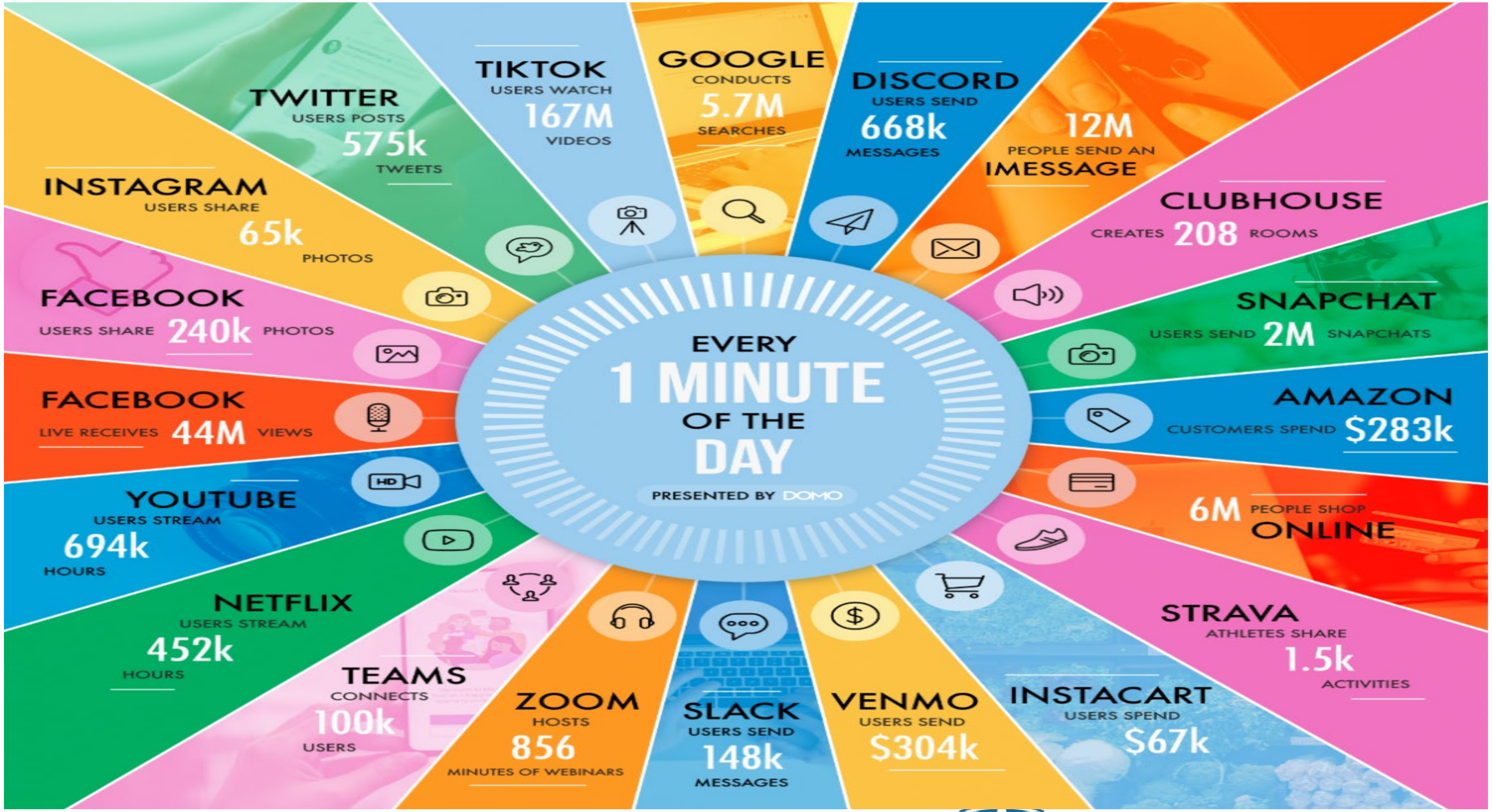
Source: https://explodingtopics.com/blog/mobile-internet-traffic

https://dailyinfographic.com/how-much-data-is-generated-every-minute

Evolution of Data

Evolution of technology

Social Media



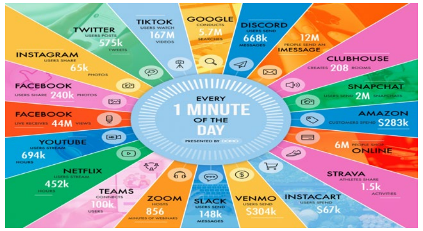
https://dailyinfographic.com/how-much-data-is-generated-every-minute

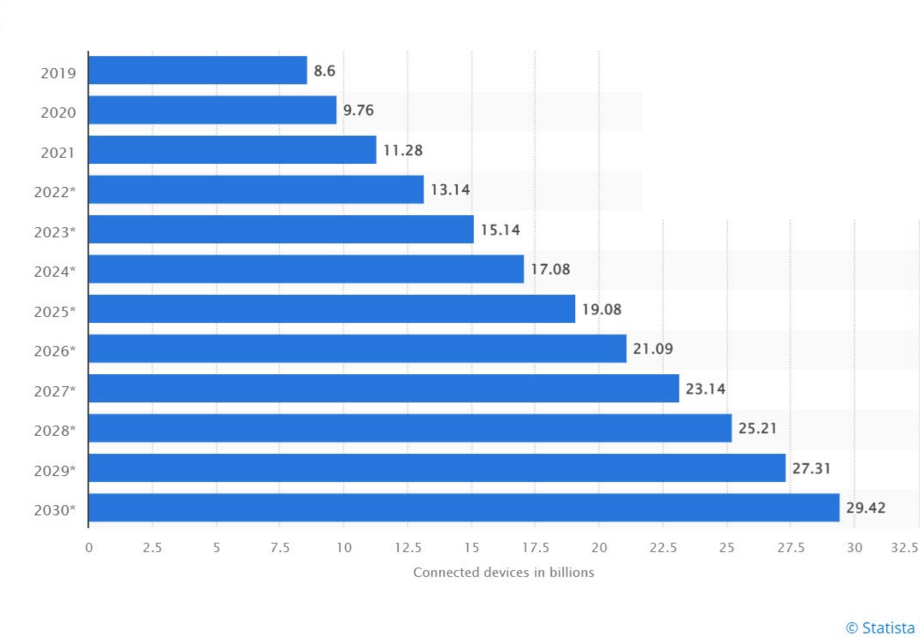
Evolution of Data

IOT

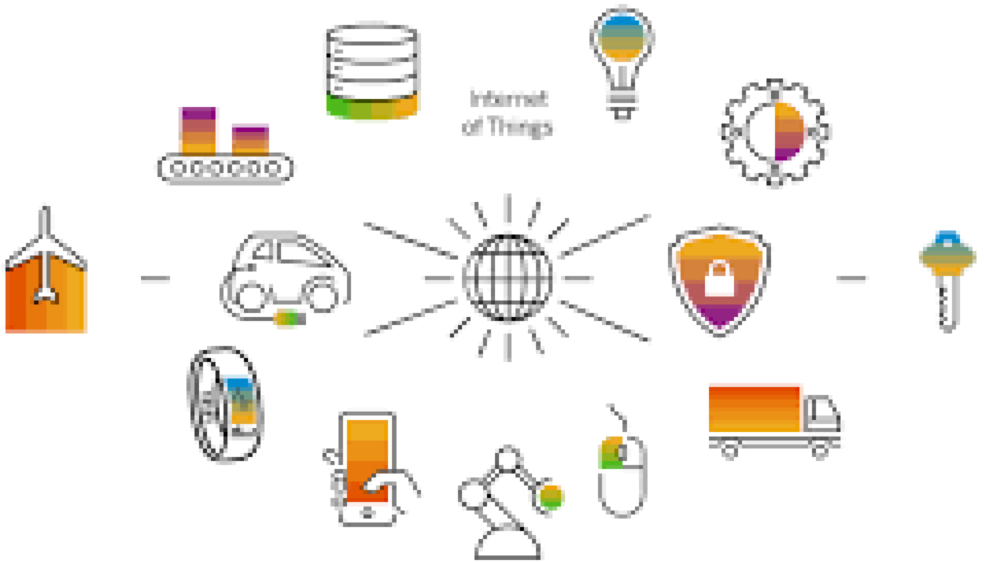
Evolution of technology

Social Media





© Statista 2023

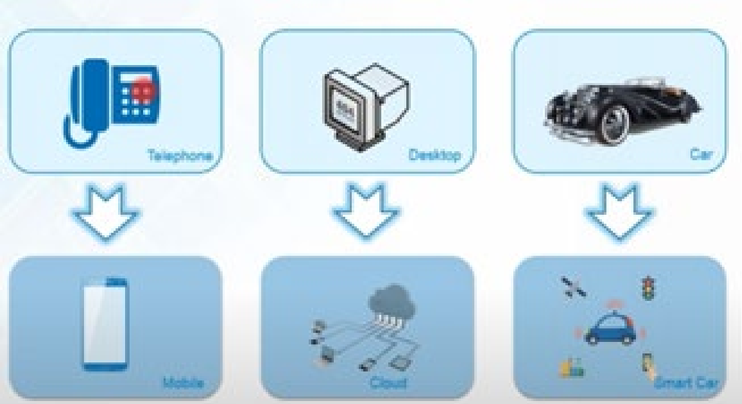


Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030

https://dailyinfographic.com/how-much-data-is-generated-every-minute

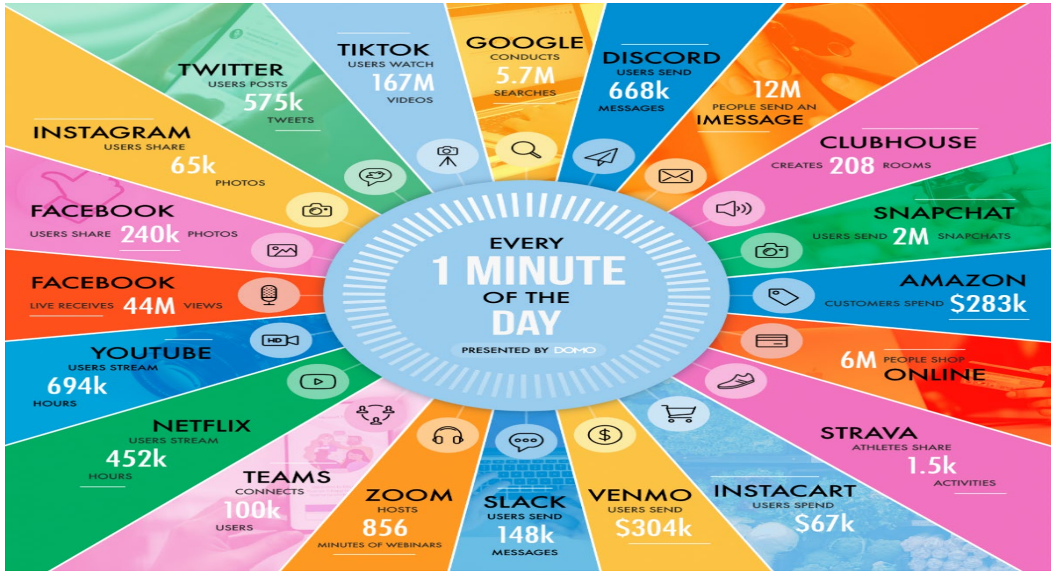
Evolution of Data

Evolution of technology



IOT

Social Media

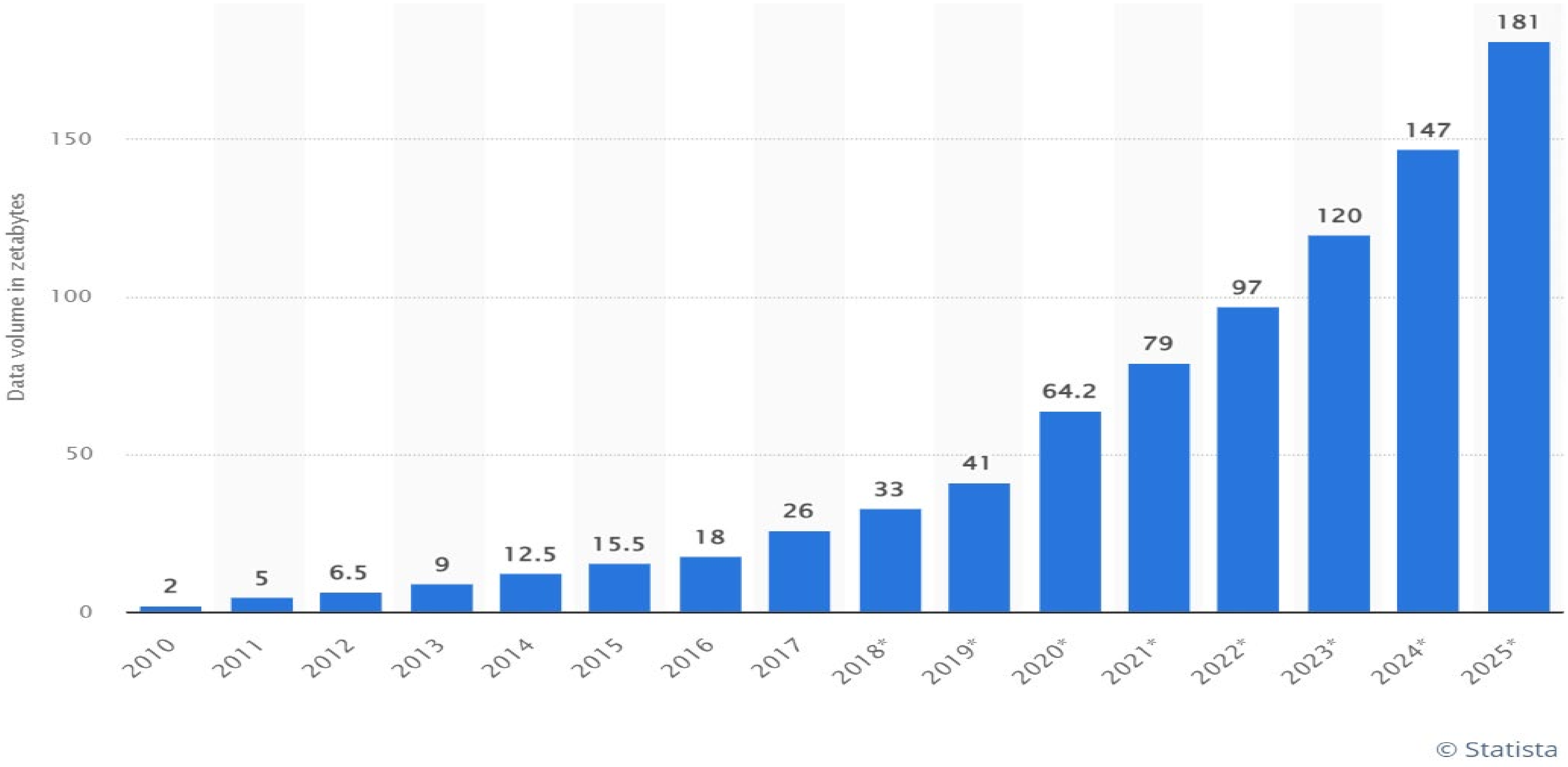


HEALTHCARE
EDUCATION
RETAIL
MEDIA
GOVERNMENT

https://www.iottechtrends.com/history-of-iot/

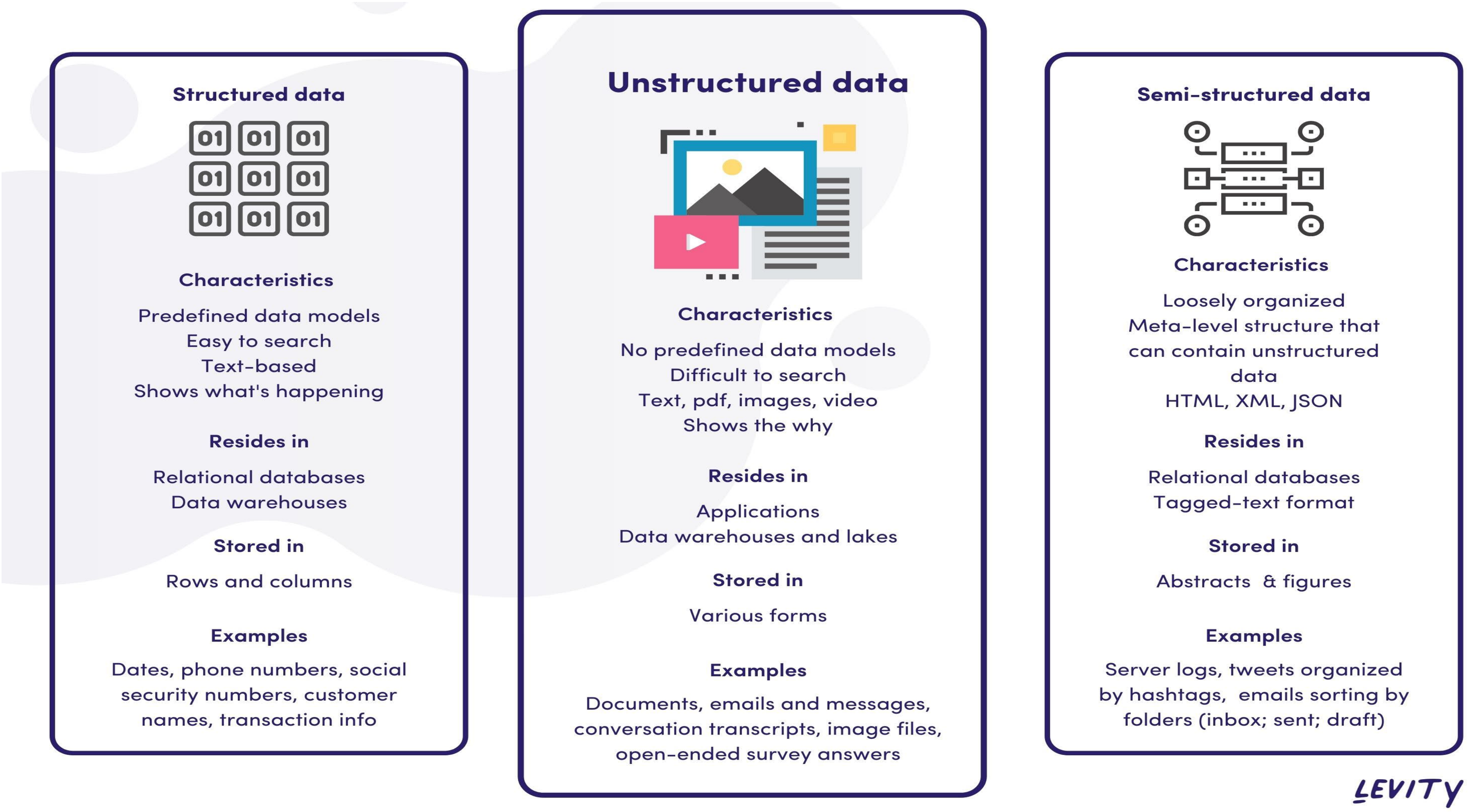
https://dailyinfographic.com/how-much-data-is-generated-every-minute

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025



https://www.statista.com/statistics/871513/worldwide-data-created/

Classification of Data



Structured data

Characteristics

Predefined data models
Easy to search
Text-based
Shows what's happening

Resides in

Relational databases
Data warehouses

Stored in

Rows and columns

Examples

Dates, phone numbers, social security numbers, customer names, transaction info

Unstructured data

Characteristics

No predefined data models
Difficult to search
Text, pdf, images, video
Shows the why

Resides in

Applications
Data warehouses and lakes

Stored in

Various forms

Examples

Documents, emails and messages, conversation transcripts, image files, open-ended survey answers

Semi-structured data

Characteristics

Loosely organized
Meta-level structure that can contain unstructured data
HTML, XML, JSON

Resides in

Relational databases
Tagged-text format

Stored in

Abstracts & figures

Examples

Server logs, tweets organized by hashtags, emails sorting by folders (inbox; sent; draft)

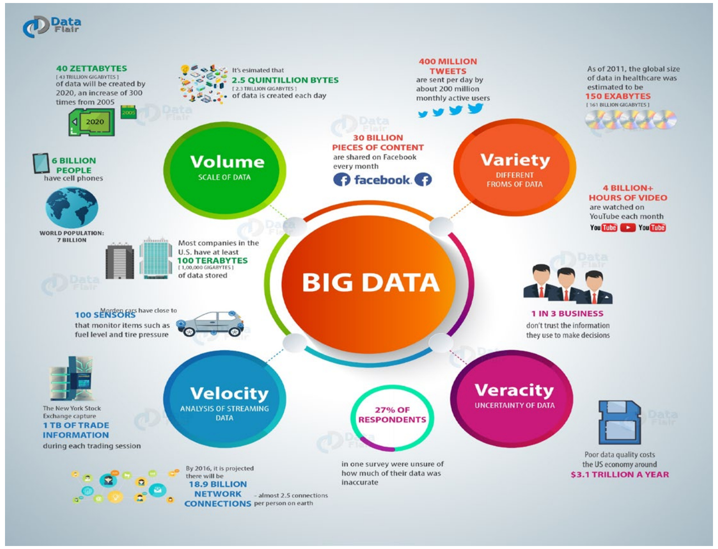
LEVITY

https://levity.ai/blog/data-classification-types-applications

Historical Interpretation of Big Data

Douglas Laney's 3Vs:
He noticed that due to surging of e-commerce activities, data has grown along three dimensions, namely:

1. **Volume**: means the incoming data stream and cumulative volume of data
2. **Velocity**: represents the pace of data used to support interaction and generated by interactions
3. **Variety**: signifies the variety of incompatible and inconsistent data formats and data structures


Data Flair
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's esimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data is created each day
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
facebook
Variety
DIFFERENT FROMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
YouTube
Most companies in the U.S. have at least
100 TERABYTES
[1,00,000 GIGABYTES]
of data stored
BIG DATA
Morden cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
1 IN 3 BUSINESS
don't trust the information they use to make decisions
The New York Stock Exchange capture
1 TB OF TRADE INFORMATION
during each trading session
Velocity
ANALYSIS OF STREAMING DATA
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
- almost 2.5 connections per person on earth

https://dzone.com/articles/why-is-big-data-in-buzz

IBM —4Vs:
IBM added another attribute or "V" for "Veracity" on the top of Douglas Laney's 3Vs notation:

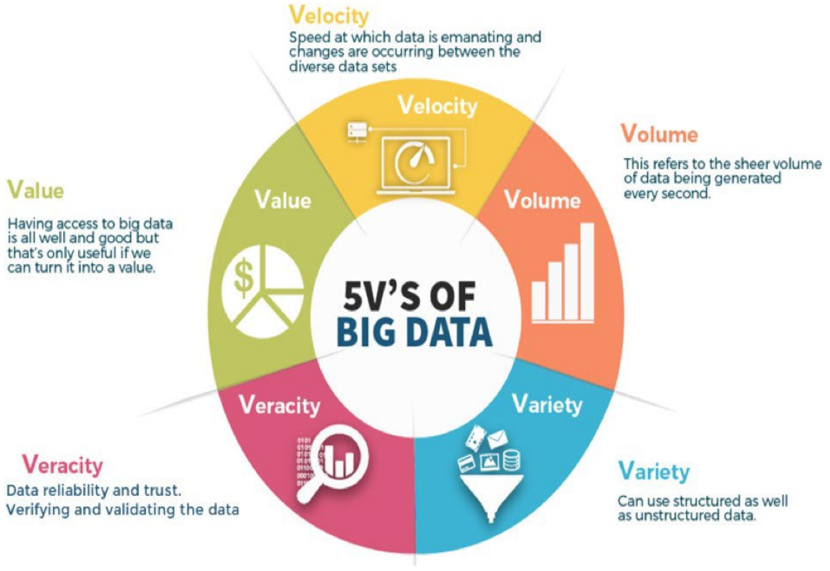
1. Volume stands for the scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. **Veracity** implies the uncertainty of data. Accuracy and trustworthiness of data is termed as veracity

Historical Interpretation Of Big Data

Yuri Demchenko’s5Vs:

Yuri added the value dimension along with the IBM 4Vs’ definition in 2013:

1. Volume stands for the scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. Veracity implies the uncertainty of data
5. **Value** refers to how useful the data is in decision making. we must be able to analyze the data to generate valuable knowledge that helps in decision making.


Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets
Volume
This refers to the sheer volume of data being generated every second.
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.
Velocity
Value
Volume
5V'S OF BIG DATA
Veracity
Variety
Veracity
Data reliability and trust. Verifying and validating the data
Variety
Can use structured as well as unstructured data.

https://www.techentice.com/the-data-veracity-big-data/

Microsoft —6Vs:

For the sake of maximizing the business value, Microsoft extended Douglas Laney’s 3Vs attributes to 6 Vs:

1. Volume stands for scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. Veracity focuses on trustworthiness of data sources

5. Variability refers to the complexity of data set. In comparison with “Variety” (or different data format), it means the number of variables in data sets.

6. Visibility emphasizes that you need to have a full picture of data in order to make informative decision

Big Data Analytics (BDA)

“Big data analytics is the often-complex process of examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions”

- **Explanatory analytics** focuses on discovering and explaining data characteristics based on existing data
- **Predictive analytics** focuses on predicting future data outcomes with a high degree of accuracy
- **Big Data Analytics-Benefits**
  - Effective marketing,
  - New revenue opportunities,
  - Customer personalization
  - Improved operational efficiency
  - Enhanced competitive edge

Over 97 percent of organizations are investing in big data and artificial intelligence.

Source: https://www.simplilearn.com/big-data-analytics-and-ai-to-manage-pandemics-article

On 16th Jan 2022, 858 Data Scientist jobs available.

Source: https://ca.indeed.com/Data-Science-Analyst-jobs?vjk=5d891b57d22a8a7b

https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Examples Of Big Data Analytics in Industries

Analytics in retail enables companies to create customer recommendations based on their purchase history, resulting in personalized shopping experiences and improved customer service. It also helps with forecasting trends and making strategic decisions based on market analysis.

Healthcare big data analytics drive quicker responses to emerging diseases and improve direct patient care, the customer experience, and administrative, insurance and payment processing.

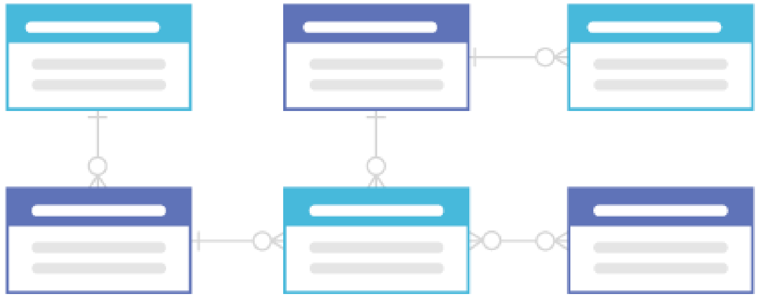
Financial analytics improve customer targeting using customer analytics. Businesses can make better informed underwriting decisions and provide better claims management while mitigating risk and fraud.

https://www.ibm.com/analytics/hadoop/big-data-analytics

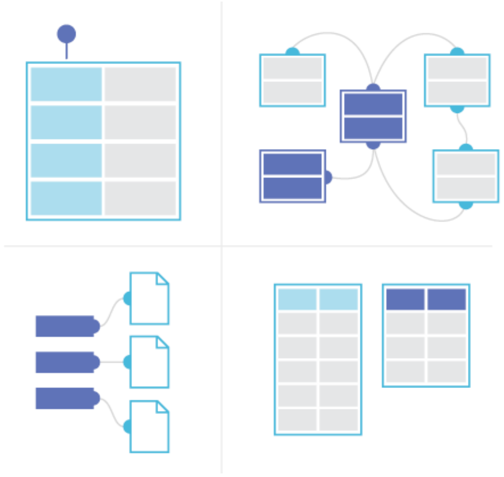
Questions Reflect the Bottom Line Of BI

1. How to store massive data (such as in PB or EB scale currently) or information in the available resources
2. How to access these massive data or information quickly
3. How to work with datasets in variety formats: structured, semi-structured, and unstructured
4. How to process these datasets in a full scalable, fault tolerant, and flexible manner
5. How to extract BI interactively and cost-effectively

Relational Data to Big Data



RDBMS



Non-Relational Database

Drawback with RDBMS for Large dataset

- Processing large data may fail
- Can’t store unstructured data
- High cost.

https://www.cognizantsoftvision.com/blog/azure/

Hadoop Ecosystem

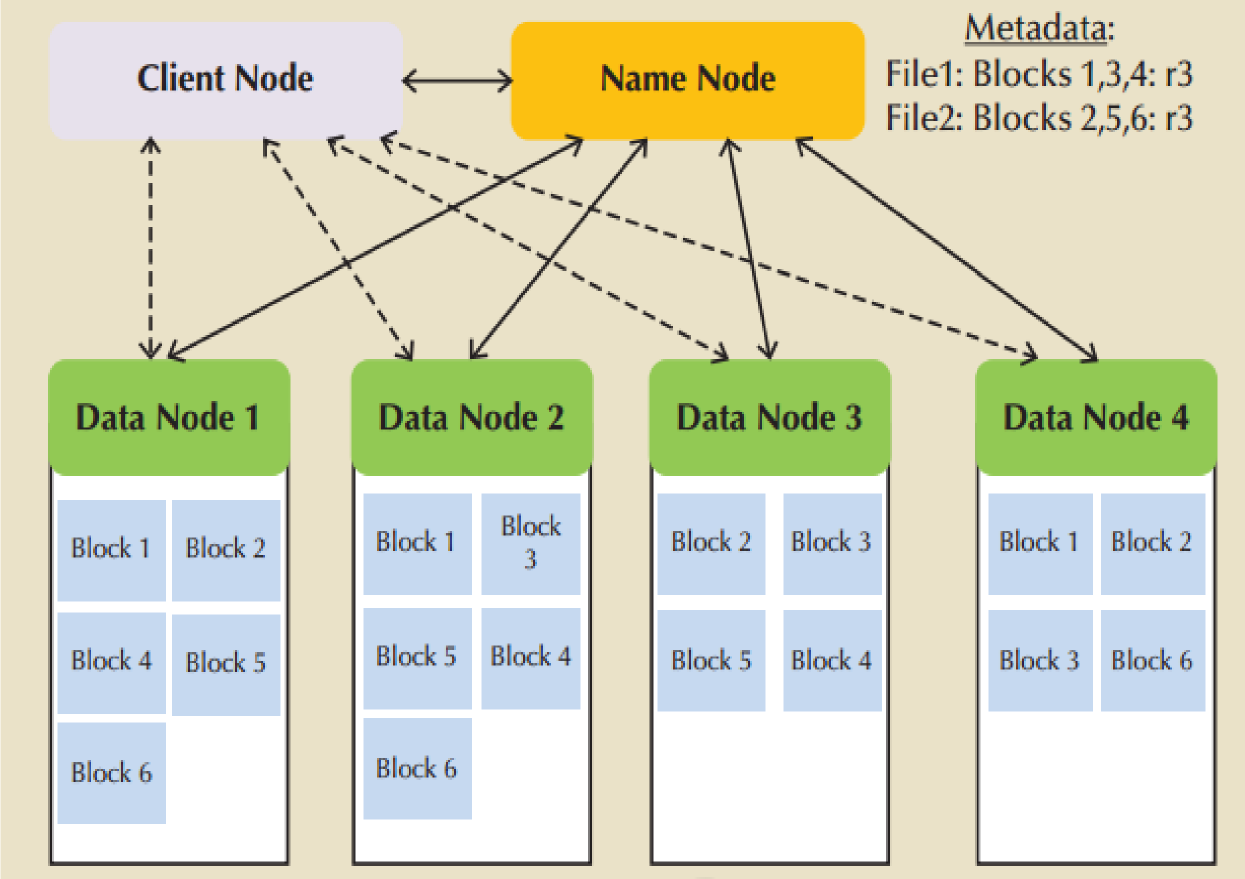
- De facto standard for most Big Data stora
- Java-based framework for distributing and processing very large data sets across clusters of computers
- Most important components:
  - **Hadoop Distributed File System (HDFS)**: Low-level distributed file processing system that can be used directly for data storage
  - **MapReduce:** Programming model that supports processing large data sets

Hadoop Distributed File System (HDFS)

- Approach based on several key assumptions:
  - ***High volume*** - Default block sizes is 64 MB and can be configured to even larger values
  - ***Write-once, read-many*** - Hadoop we can store all kinds of data once which can be accessed any number of times. Model simplifies concurrent issues and improves data throughput
  - ***Streaming access*** - Hadoop is optimized for batch processing of entire files as a continuous stream of data
  - ***Fault tolerance*** – HDFS is designed to replicate data across many different devices so that when one fails, data is still available from another device

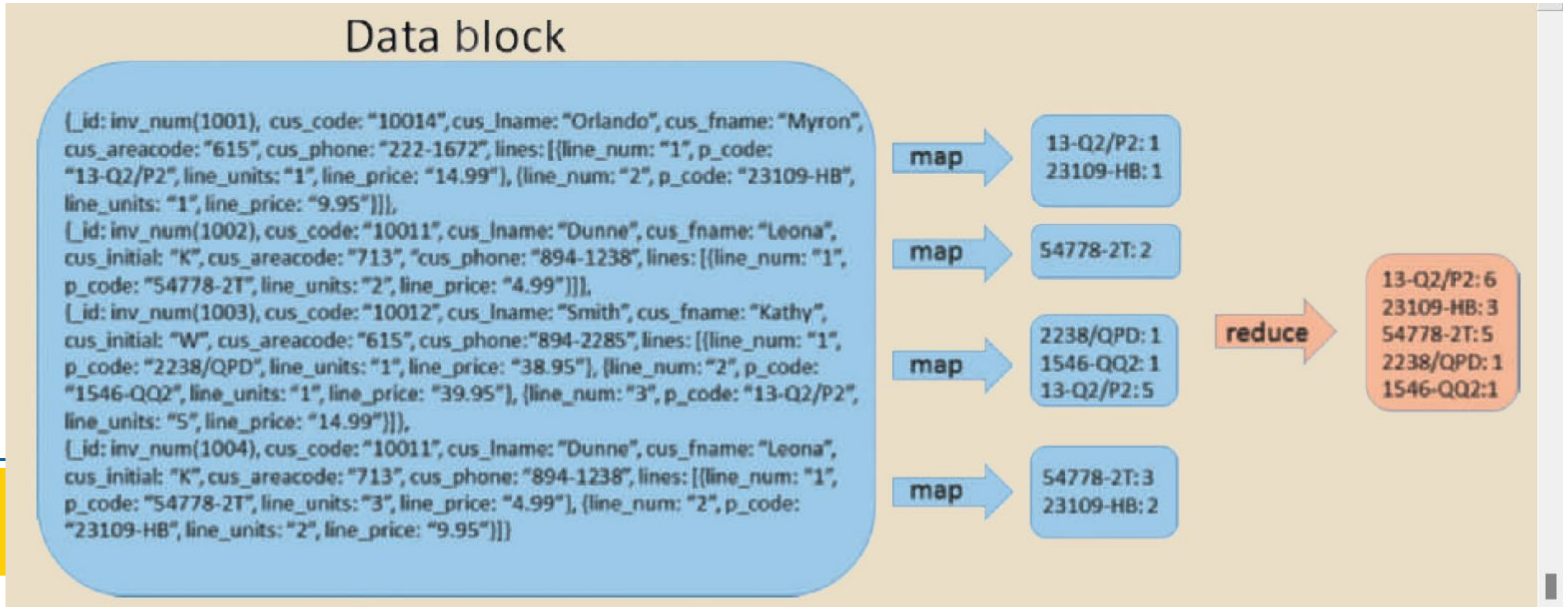
Hadoop Distributed File System (HDFS)

- Uses several types of nodes (computers):
  - Data node store the actual file data
  - Name node contains file system metadata
  - Client node makes requests to the file system as needed to support user applications



MapReduce

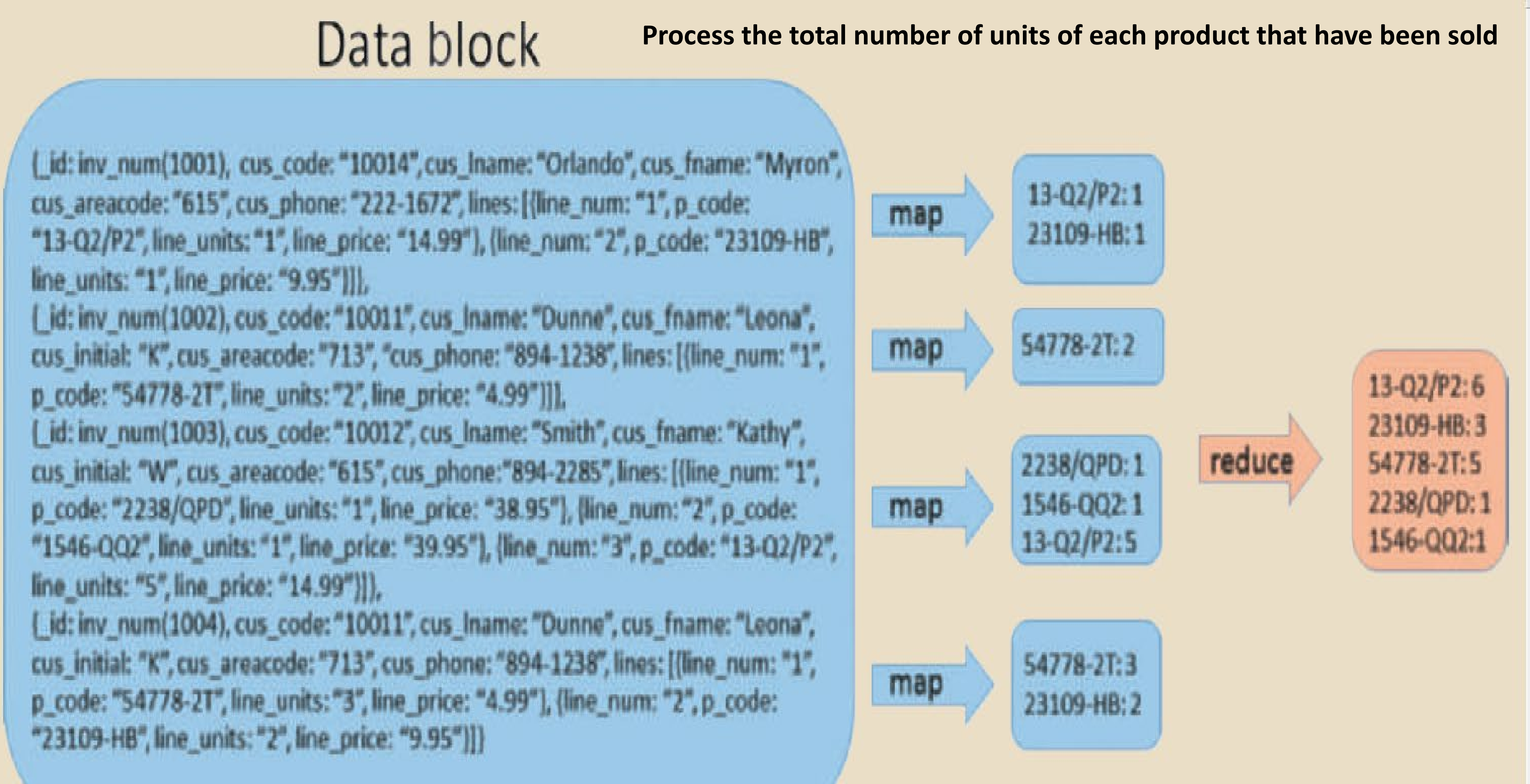
- Framework used to process large data sets across clusters
  - Breaks down complex tasks into smaller subtasks, performing the subtasks and producing a final result
  - **Map** function takes a collection of data and sorts and filters it into a set of key-value pairs
    - **Mapper** program performs the map function
  - **Reduce** summaries results of map function to produce a single result
    - **Reducer** program performs the reduce function



Data block

```
(_id: inv_num(1001), cus_code: "10014",cus_lname: "Orlando", cus_fname: "Myron",
cus_areacode: "615", cus_phone: "222-1672", lines: [{line_num: "1", p_code:
"13-Q2/P2", line_units: "1", line_price: "14.99"}, {line_num: "2", p_code: "23109-HB",
line_units: "1", line_price: "9.95"}]},
(_id: inv_num(1002), cus_code: "10011", cus_lname: "Dunne", cus_fname: "Leona",
cus_initial: "K", cus_areacode: "713", "cus_phone: "894-1238", lines: [{line_num: "1",
p_code: "54778-2T", line_units: "2", line_price: "4.99"}]],
(_id: inv_num(1003), cus_code: "10012", cus_lname: "Smith", cus_fname: "Kathy",
cus_initial: "W", cus_areacode: "615", cus_phone:"894-2285", lines: [{line_num: "1",
p_code: "2238/QPD", line_units: "1", line_price: "38.95"}, {line_num: "2", p_code:
"1546-QQ2", line_units: "1", line_price: "39.95"}, {line_num: "3", p_code: "13-Q2/P2",
line_units: "5", line_price: "14.99"}]},
(_id: inv_num(1004), cus_code: "10011", cus_lname: "Dunne", cus_fname: "Leona",
cus_initial: "K", cus_areacode: "713", cus_phone: "894-1238", lines: [{line_num: "1",
p_code: "54778-2T", line_units: "3", line_price: "4.99"}, {line_num: "2", p_code:
"23109-HB", line_units: "2", line_price: "9.95"}]}
```

MapReduce



Data block

Process the total number of units of each product that have been sold

{_id: inv_num(1001), cus_code: “10014”,cus_lname: “Orlando”, cus_fname: “Myron”, cus_areacode: “615”, cus_phone: “222-1672”, lines: [{line_num: “1”, p_code: “13-Q2/P2”, line_units: “1”, line_price: “14.99”}, {line_num: “2”, p_code: “23109-HB”, line_units: “1”, line_price: “9.95”}]},

{_id: inv_num(1002), cus_code: “10011”, cus_lname: “Dunne”, cus_fname: “Leona”, cus_initial: “K”, cus_areacode: “713”, “cus_phone: “894-1238”, lines: [{line_num: “1”, p_code: “54778-2T”, line_units: “2”, line_price: “4.99”}]},

{_id: inv_num(1003), cus_code: “10012”, cus_lname: “Smith”, cus_fname: “Kathy”, cus_initial: “W”, cus_areacode: “615”, cus_phone: “894-2285”, lines: [{line_num: “1”, p_code: “2238/QPD”, line_units: “1”, line_price: “38.95”}, {line_num: “2”, p_code: “1546-QQ2”, line_units: “1”, line_price: “39.95”}, {line_num: “3”, p_code: “13-Q2/P2”, line_units: “5”, line_price: “14.99”}]},

{_id: inv_num(1004), cus_code: “10011”, cus_lname: “Dunne”, cus_fname: “Leona”, cus_initial: “K”, cus_areacode: “713”, cus_phone: “894-1238”, lines: [{line_num: “1”, p_code: “54778-2T”, line_units: “3”, line_price: “4.99”}, {line_num: “2”, p_code: “23109-HB”, line_units: “2”, line_price: “9.95”}]}

13-Q2/P2: 1
23109-HB: 1

54778-2T: 2

2238/QPD: 1
1546-QQ2: 1
13-Q2/P2: 5

54778-2T: 3
23109-HB: 2

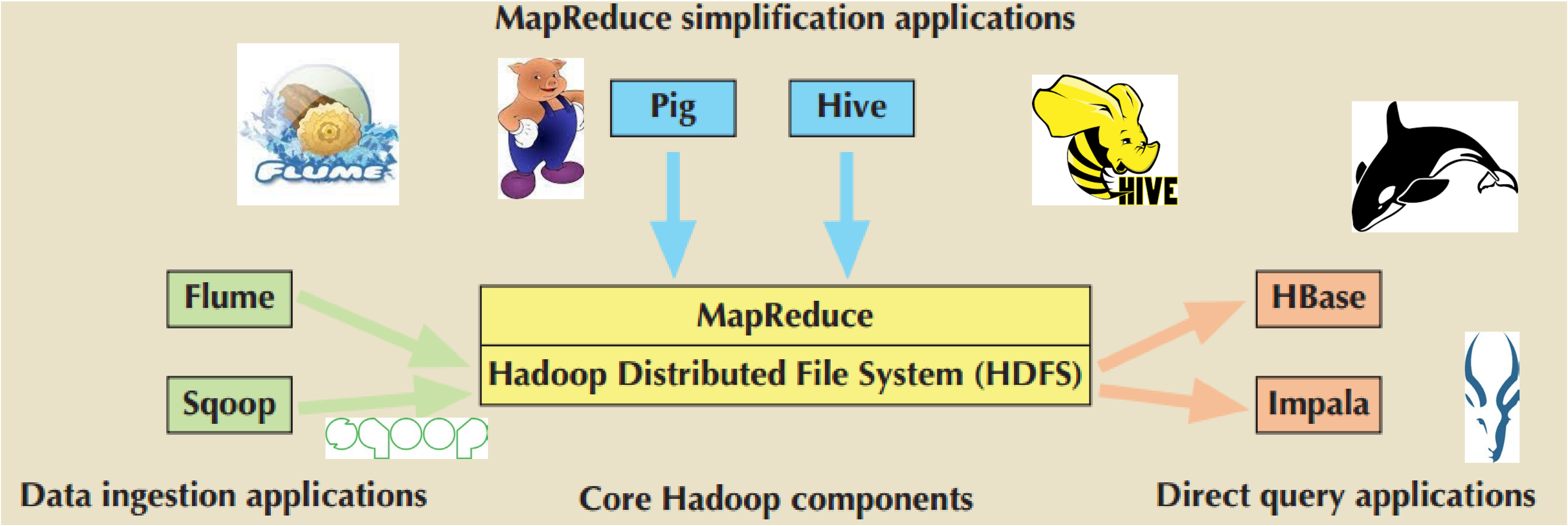
13-Q2/P2: 6
23109-HB: 3
54778-2T: 5
2238/QPD: 1
1546-QQ2: 1

A Sample of the Hadoop Ecosystem

Map Reduce Simplification Applications:

Hive is a data warehousing system that sites on top of HDFS and supports its own SQL-like language

Pig compiles a high-level scripting language (Pig Latin) into MapReduce jobs for executing in Hadoop


MapReduce simplification applications
FLUME
Pig
Hive
HIVE
Flume
Sqoop
sqoop
MapReduce
Hadoop Distributed File System (HDFS)
HBase
Impala
Data ingestion applications
Core Hadoop components
Direct query applications

Data Ingestion Applications:

Flume is a component for ingesting data in Hadoop

Sqoop is a tool for converting data back and forth between a relational database and the HDFS

Direct Query Applications:

HBase is a column-oriented NoSQL database designed to sit on top of the HDFS that quickly processes sparse datasets

Impala was the first SQL-on-Hadoop application

SPARK

Spark is a unified analytics engine for large-scale data processing. Spark is a fast- and general-purpose computation platform based on large clusters.

It was developed by the UC Berkeley RAD Lab (now called as AMP Lab).

Speed: Uses In-Memory Processing- 100x faster than disk access

Ease of Use: Applications can be written in Java, Scala,Python, R etc.

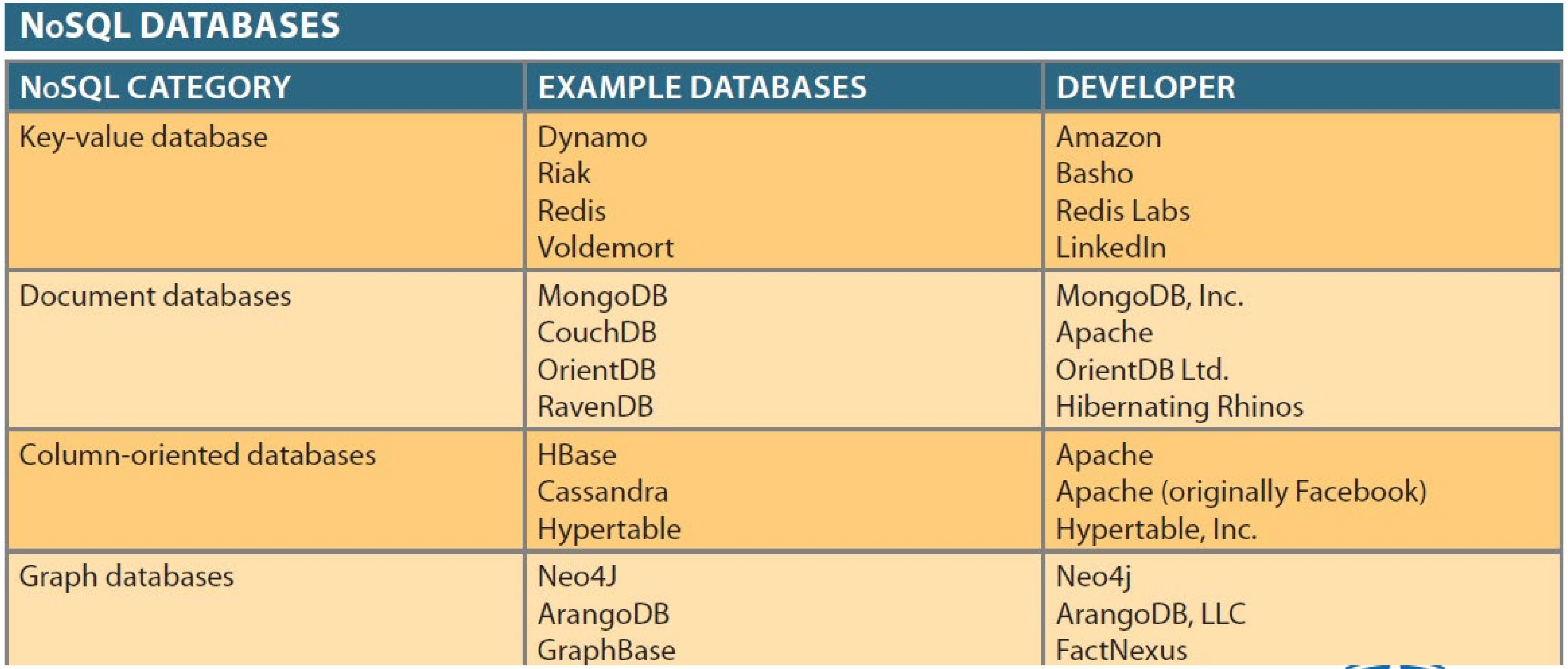
Generality: Can combine SQL, Streaming and Complex Analytics

Access Diverse Data Stores: Hadoop, Apache Mesos(Cluster Management) , Kubernetes (Application Deployment) etc.

Open source: Apache Spark

NoSQL

- Name given to non-relational database technologies developed to address Big data challenges.
- There are literally hundreds of products that can be considered as being under the broadly defined term NoSQL.
- Most of these fit roughly into one of four categories:

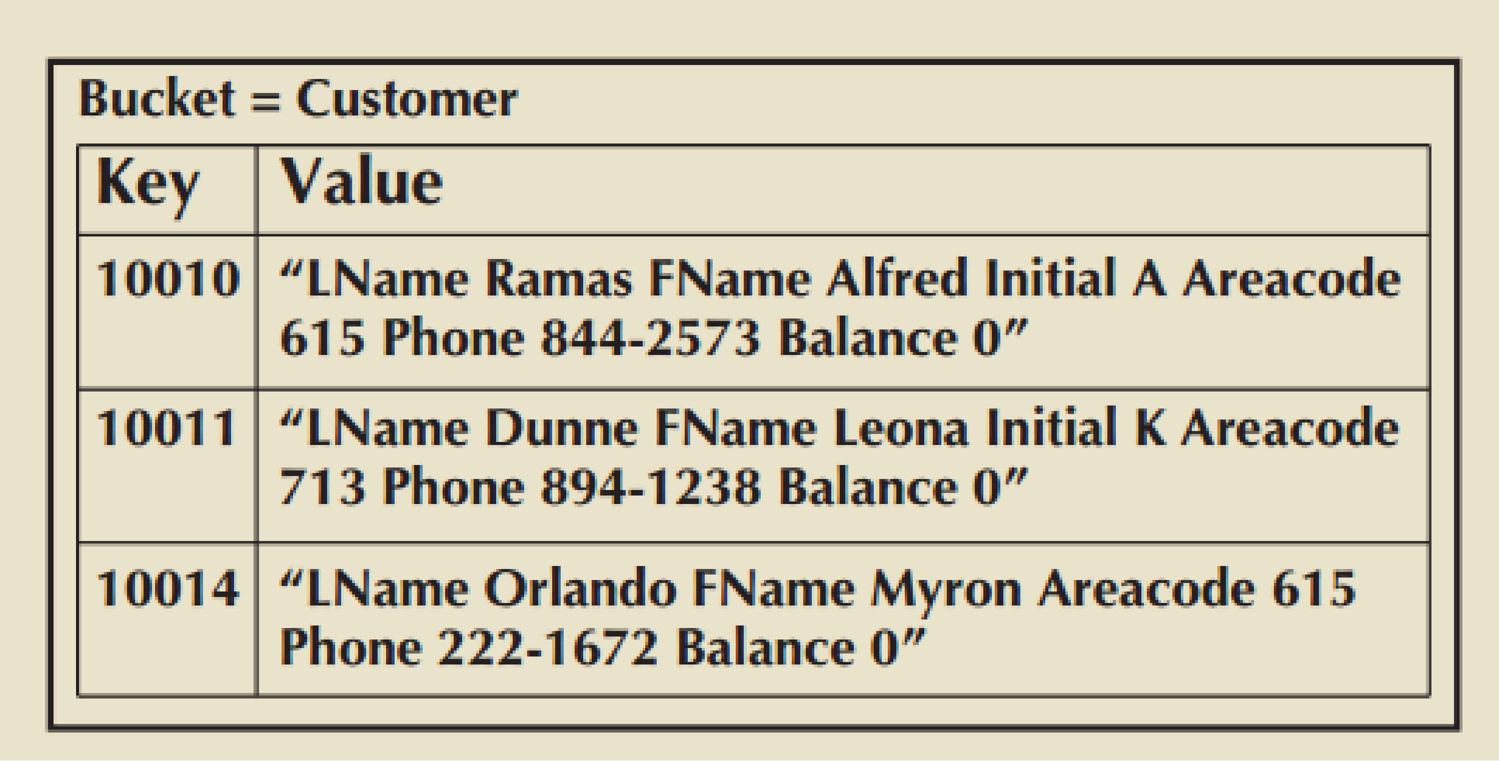


NoSQL DATABASES

NoSQL CATEGORY	EXAMPLE DATABASES	DEVELOPER
Key-value database	Dynamo Riak Redis Voldemort	Amazon Basho Redis Labs LinkedIn
Document databases	MongoDB CouchDB OrientDB RavenDB	MongoDB, Inc. Apache OrientDB Ltd. Hibernating Rhinos
Column-oriented databases	HBase Cassandra Hypertable	Apache Apache (originally Facebook) Hypertable, Inc.
Graph databases	Neo4J ArangoDB GraphBase	Neo4j ArangoDB, LLC FactNexus

NoSQL

Key-value (KV) databases store data as a collection of key-value pairs organized as **buckets** which are the equivalent of tables



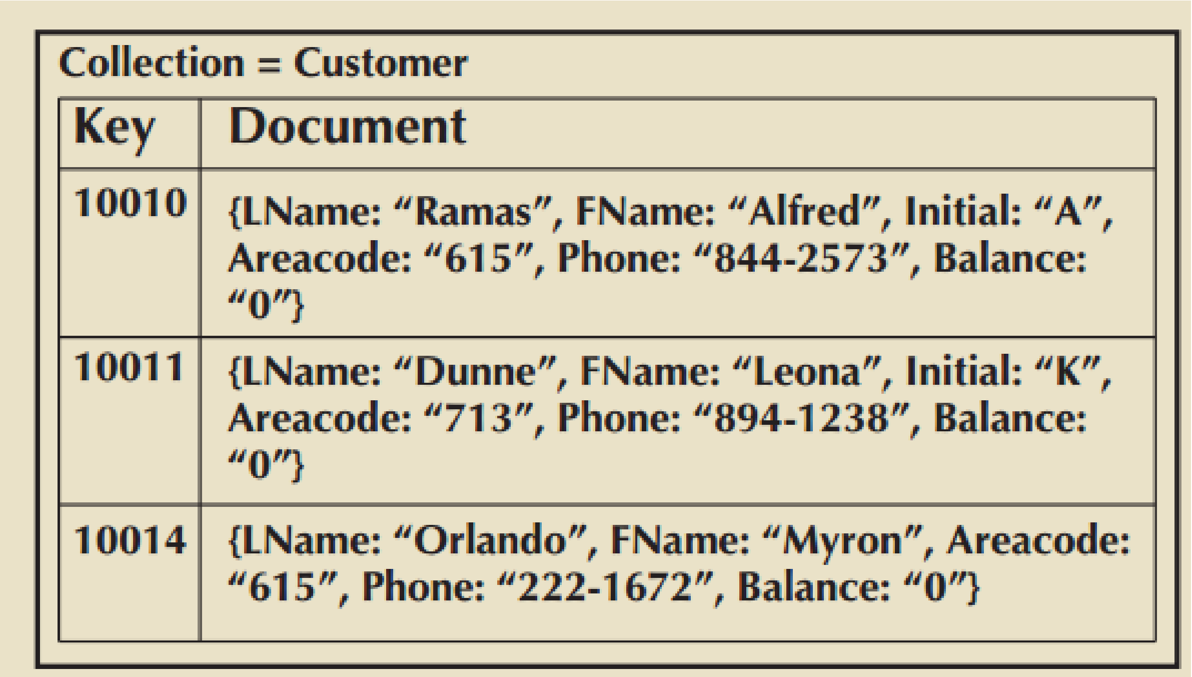
Bucket = Customer

Key	Value
10010	"LName Ramas FName Alfred Initial A Areacode 615 Phone 844-2573 Balance 0"
10011	"LName Dunne FName Leona Initial K Areacode 713 Phone 894-1238 Balance 0"
10014	"LName Orlando FName Myron Areacode 615 Phone 222-1672 Balance 0"

Ex: Redis, Oracle NoSQL

NoSQL

Document databases store data in key-value pairs in which the value components are tag-encoded documents grouped into logical groups called **collections**



Collection = Customer

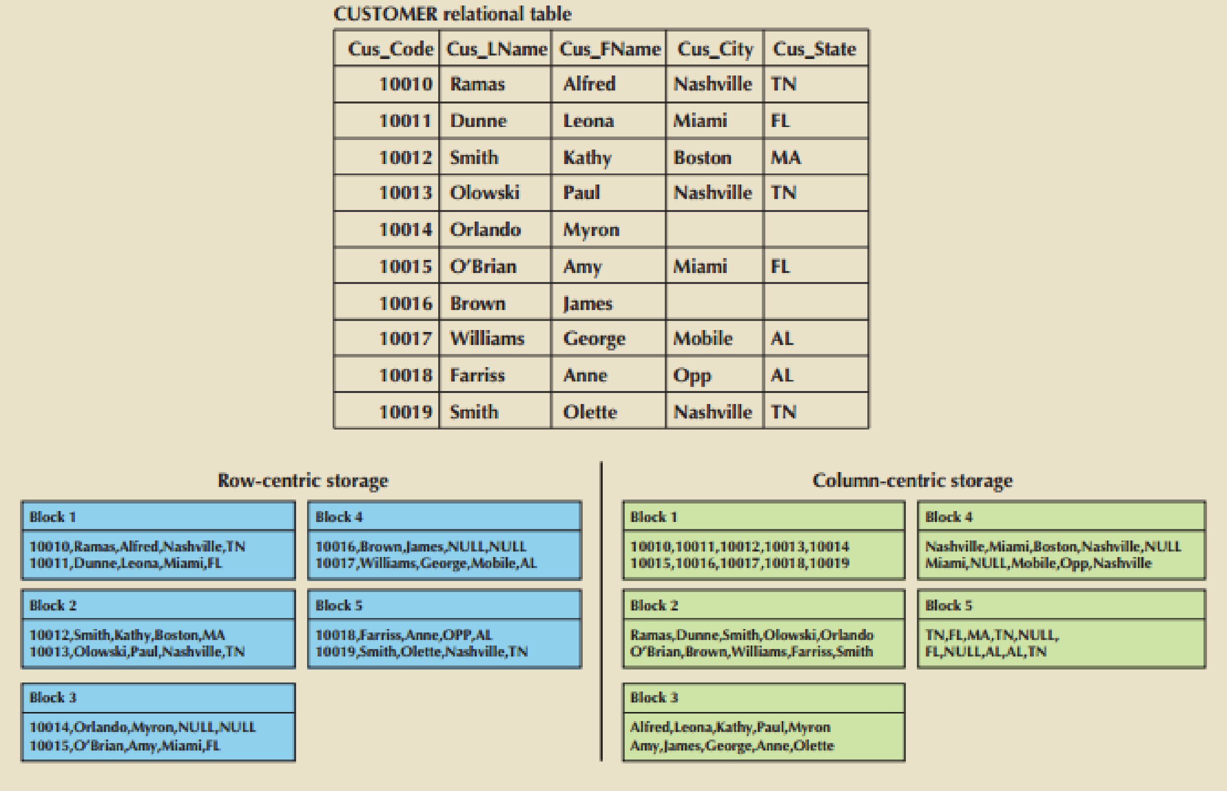
Key	Document
10010	{LName: “Ramas”, FName: “Alfred”, Initial: “A”, Areacode: “615”, Phone: “844-2573”, Balance: “0”}
10011	{LName: “Dunne”, FName: “Leona”, Initial: “K”, Areacode: “713”, Phone: “894-1238”, Balance: “0”}
10014	{LName: “Orlando”, FName: “Myron”, Areacode: “615”, Phone: “222-1672”, Balance: “0”}

Ex: MongoDB

NoSQL

Column-oriented databases refers to two technologies:

- **Column-centric storage**: Data stored in blocks which hold data from a single column across many rows
- **Row-centric storage:** Data stored in block which hold data from all columns of a given set of rows



CUSTOMER relational table

Cus_Code	Cus_LName	Cus_FName	Cus_City	Cus_State
10010	Ramas	Alfred	Nashville	TN
10011	Dunne	Leona	Miami	FL
10012	Smith	Kathy	Boston	MA
10013	Olowski	Paul	Nashville	TN
10014	Orlando	Myron		
10015	O'Brian	Amy	Miami	FL
10016	Brown	James		
10017	Williams	George	Mobile	AL
10018	Farriss	Anne	Opp	AL
10019	Smith	Olette	Nashville	TN

Row-centric storage

Block 1
10010,Ramas,Alfred,Nashville,TN
10011,Dunne,Leona,Miami,FL

Block 2
10012,Smith,Kathy,Boston,MA
10013,Olowski,Paul,Nashville,TN

Block 3
10014,Orlando,Myron,NULL,NULL
10015,O'Brian,Amy,Miami,FL

Block 4
10016,Brown,James,NULL,NULL
10017,Williams,George,Mobile,AL

Block 5
10018,Farriss,Anne,OPP,AL
10019,Smith,Olette,Nashville,TN

Column-centric storage

Block 1
10010,10011,10012,10013,10014
10015,10016,10017,10018,10019

Block 2
Ramas,Dunne,Smith,Olowski,Orlando
O'Brian,Brown,Williams,Farriss,Smith

Block 3
Alfred,Leona,Kathy,Paul,Myron
Amy,James,George,Anne,Olette

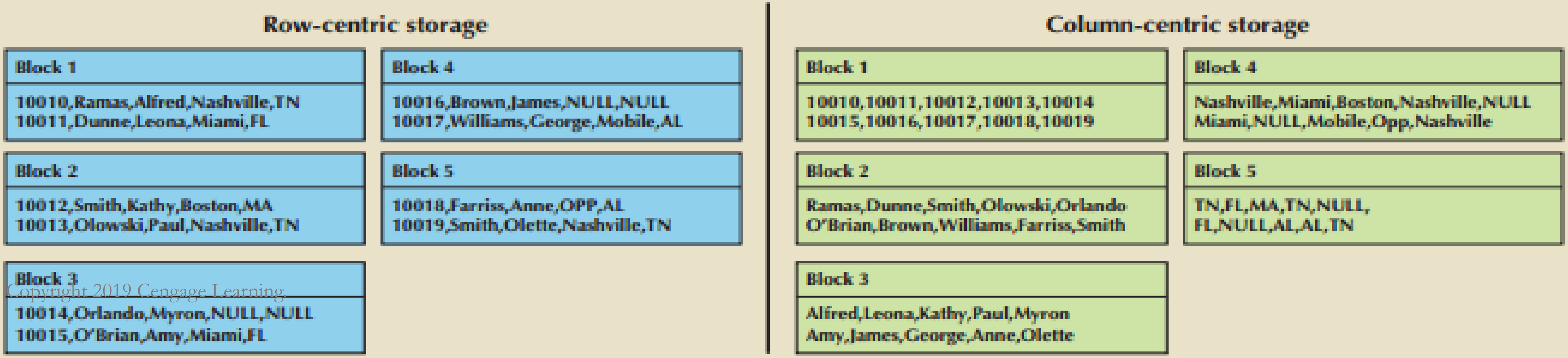
Block 4
Nashville,Miami,Boston,Nashville,NULL
Miami,NULL,Mobile,Opp,Nashville

Block 5
TN,FL,MA,TN,NULL,
FL,NULL,AL,AL,TN

Ex: Google's BigTable, HBase, and Cassandra.

CUSTOMER relational table

Cus_Code	Cus_LName	Cus_FName	Cus_City	Cus_State
10010	Ramas	Alfred	Nashville	TN
10011	Dunne	Leona	Miami	FL
10012	Smith	Kathy	Boston	MA
10013	Olowski	Paul	Nashville	TN
10014	Orlando	Myron		
10015	O'Brian	Amy	Miami	FL
10016	Brown	James		
10017	Williams	George	Mobile	AL
10018	Farriss	Anne	Opp	AL
10019	Smith	Olette	Nashville	TN



Row-centric storage

Block 1
10010,Ramas,Alfred,Nashville,TN
10011,Dunne,Leona,Miami,FL

Block 2
10012,Smith,Kathy,Boston,MA
10013,Olowski,Paul,Nashville,TN

Block 3
10014,Orlando,Myron,NULL,NULL
10015,O'Brian,Amy,Miami,FL

Block 4
10016,Brown,James,NULL,NULL
10017,Williams,George,Mobile,AL

Block 5
10018,Farriss,Anne,OPP,AL
10019,Smith,Olette,Nashville,TN

Column-centric storage

Block 1
10010,10011,10012,10013,10014
10015,10016,10017,10018,10019

Block 2
Ramas,Dunne,Smith,Olowski,Orlando
O'Brian,Brown,Williams,Farriss,Smith

Block 3
Alfred,Leona,Kathy,Paul,Myron
Amy,James,George,Anne,Olette

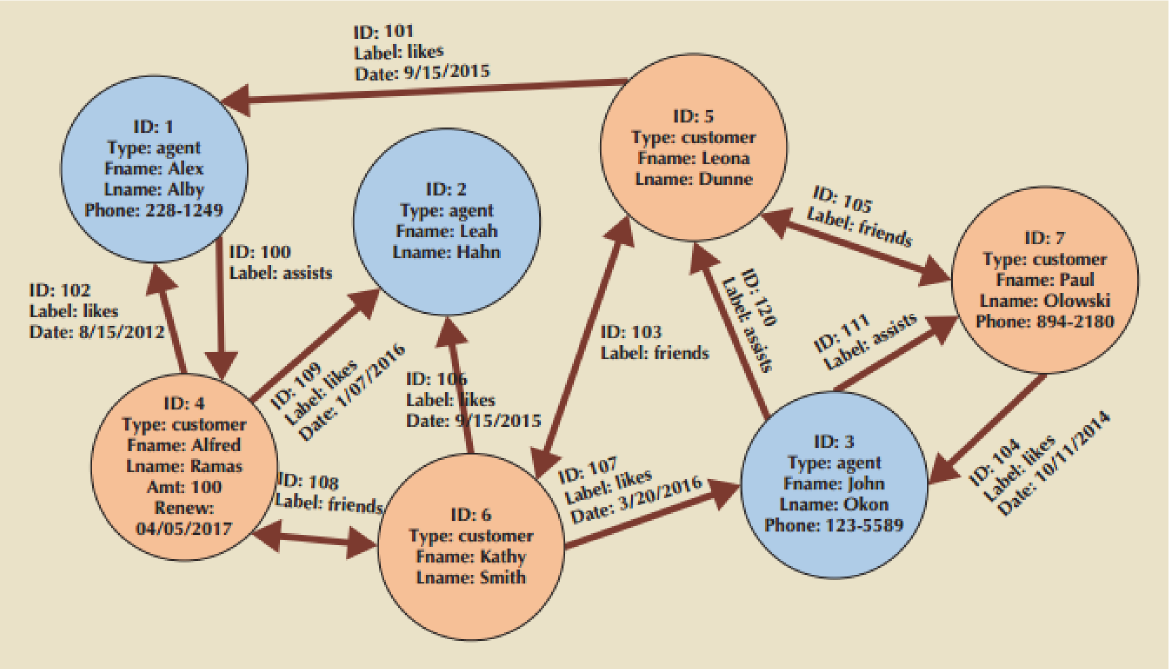
Block 4
Nashville,Miami,Boston,Nashville,NULL
Miami,NULL,Mobile,Opp,Nashville

Block 5
TN,FL,MA,TN,NULL,
FL,NULL,AL,AL,TN

NoSQL

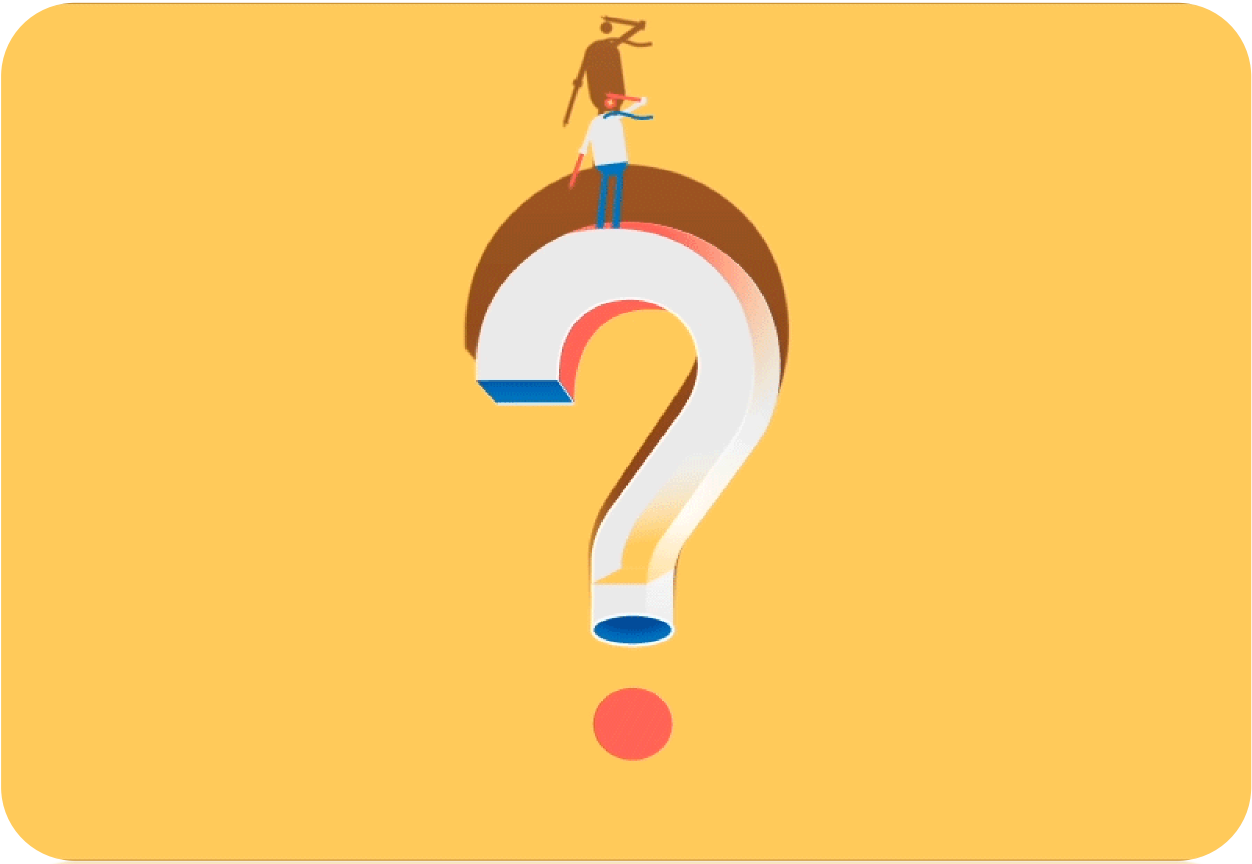
Graph databases store data on relationship-rich data as a collection of **nodes** and **edges** (relationships)

- **Properties** are the attributes of a node or edge of interest to a user
- **Traversal** is a query in a graph database



Ex: Oracle RDF (Resources Description Framework)

Any Questions



Recap and Conclusion

- Evolution of Data
- Relational Data to Big Data
- Debates of Big Data Implication
- Historical Interpretation Of Big Data
- Big Data Analytics (BDA)
- Hadoop (HDFS & Map Reduce) help to handle Big data efficiently.
- Spark a powerful open-source unified analytics engine

APPENDIX

JavaScript Object Notation (JSON)

- JSON Syntax Rules
  - Data is in name/value pairs
  - Data is separated by commas
  - Curly braces hold objects
  - Square brackets hold arrays

-
  ```
  {
  "employees":[
      {"firstName":"John", "lastName":"Doe"},
      {"firstName":"Anna", "lastName":"Smith"},
      {"firstName":"Peter", "lastName":"Jones"}
  ]
  }
  ```