BIG DATA AND NOSQL

Lecture 12

Contents

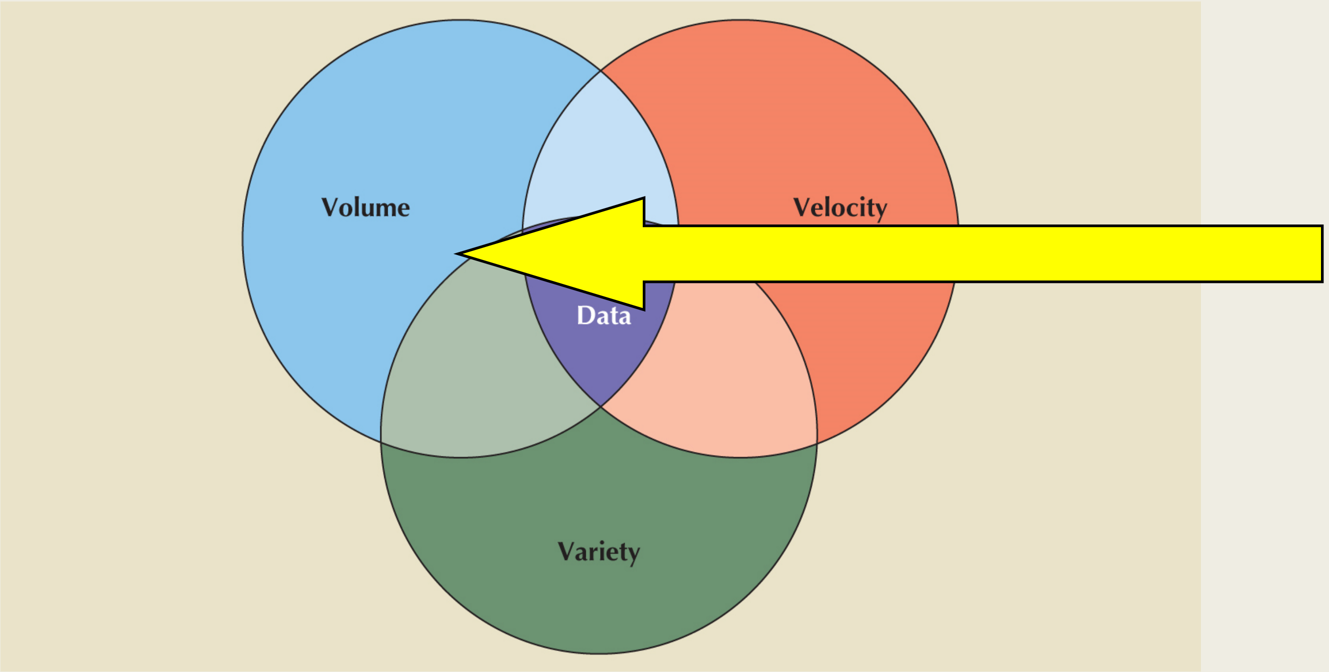
1. Big Data
2. Hadoop
3. NoSQL Databases

What is Big Data?

1. Volume
   - *quantity of data to be stored*
2. Velocity
   - *speed at which data is entered into system and must be processed*
3. Variety
   - *variations in the structure of data to be stored*

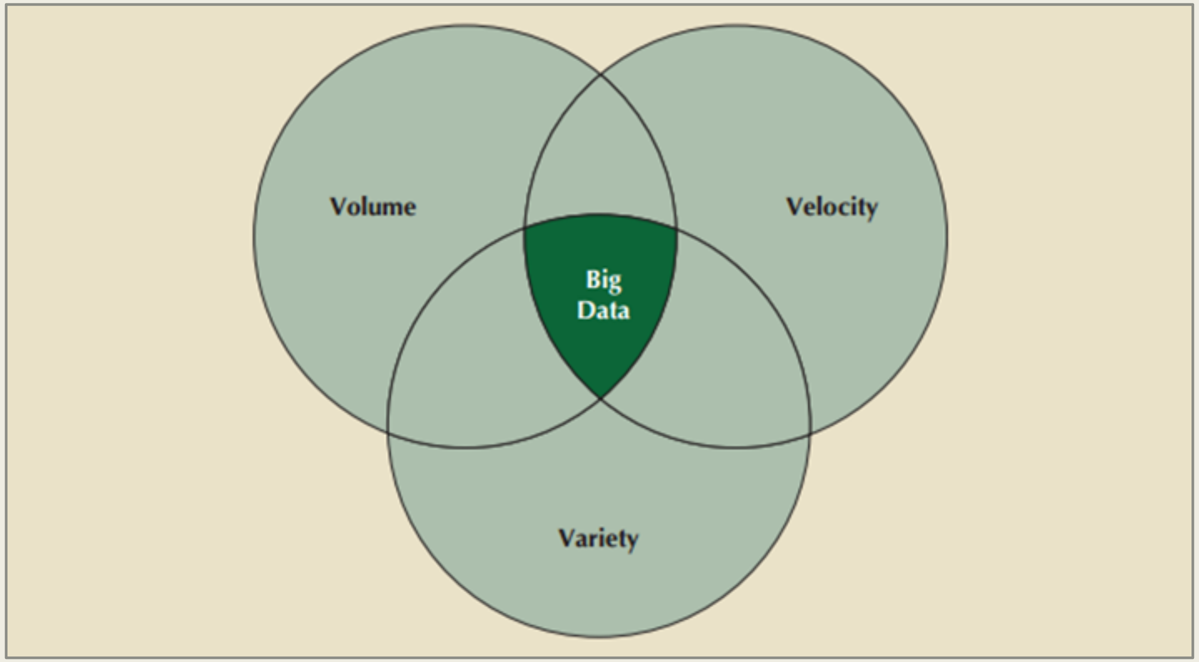
Original View of Big Data

- Originally, all of the 3 Vs required.
- For example, huge volume, high data speed, variation in data structure.



Current View of Big Data

- Currently, any of the 3 Vs, not all 3 Vs
- For example, you might have only huge volume



Big Data - Volume

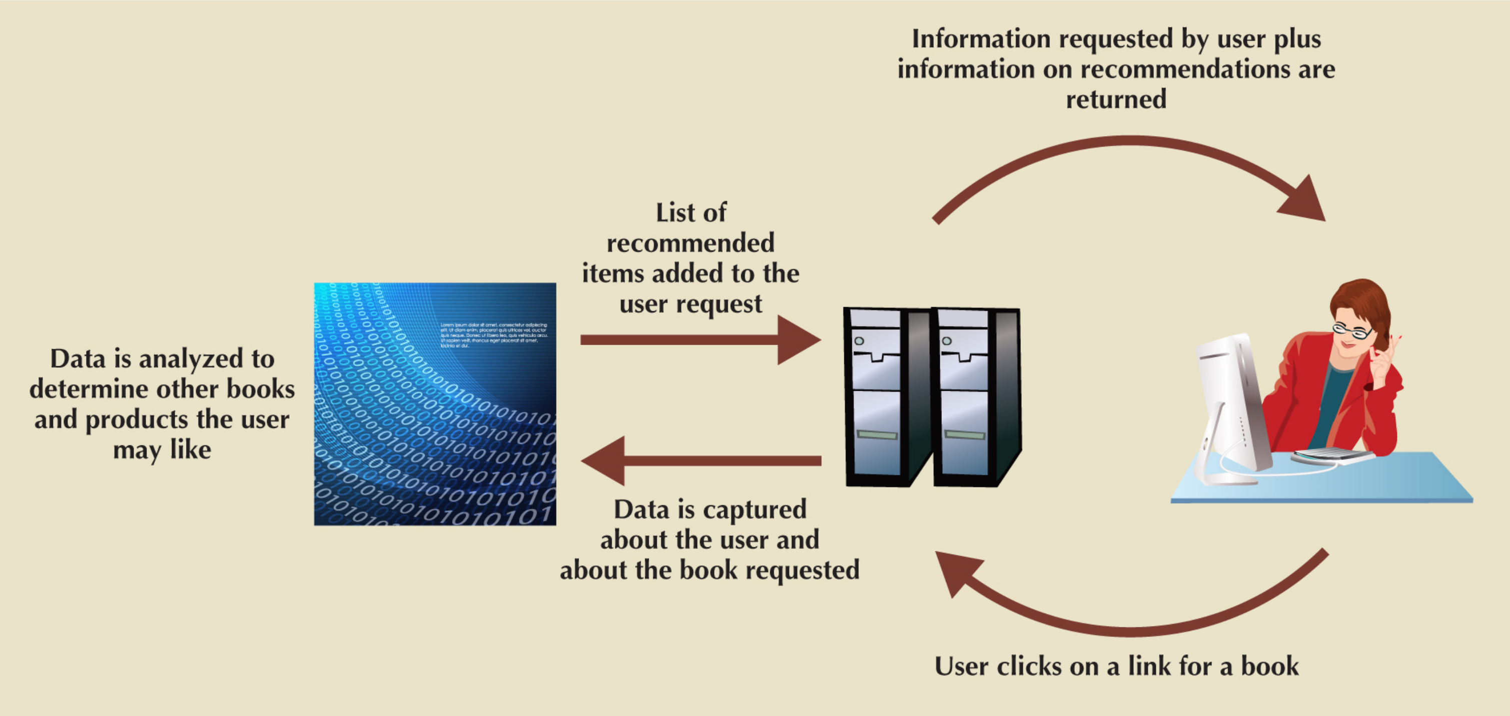
- Quantity of data to be stored
- Scaling out
- when the workload exceeds server capacity, it is spread out across a number of servers
- Scaling up
- the number of systems stays the same but migrate each to a larger system

Big Data - Velocity

- Speed at which data is entered into system and must be processed
- Stream processing
  - data stream is processed as it enters the system
- Feedback loop processing
  - analyse data to produce actionable results

Big Data - Velocity

- Feedback loop processing


Information requested by user plus information on recommendations are returned
List of recommended items added to the user request
Data is analyzed to determine other books and products the user may like
Data is captured about the user and about the book requested
User clicks on a link for a book

Big Data - Variety

- Variations in the structure of data to be stored
- Structured data
  - fits into a predefined data model, e.g., ER
- Unstructured data
  - does not fit into a predefined model

Big Data – Other Vs

- Variability
  - the meaning of data differs based on context
- Veracity
  - data trustworthiness
- Value
  - degree data can be analysed for meaningful insight
- Visualization
  - graphically present data to make it understandable

Hadoop

- https://www.youtube.com/watch?v=4DgTLaFNQq0
- https://www.youtube.com/watch?v=bcjSe0xCHbE
- https://www.youtube.com/watch?v=9s-vSeWej1U
- https://www.youtube.com/watch?v=MfF750YVDxM

Hadoop

- De facto standard for most Big Data storage and processing using commodity components
- Storage
  - Hadoop Distributed File System (HDFS)
- Processing
  - MapReduce

Hadoop - HDFS

- High Volume
  - Minimum block size is 64 MB
- Write-once, Read-many
  - Simplifies concurrent issues and improves data throughput
- Streaming access
  - Entire files processed as a continuous stream of data

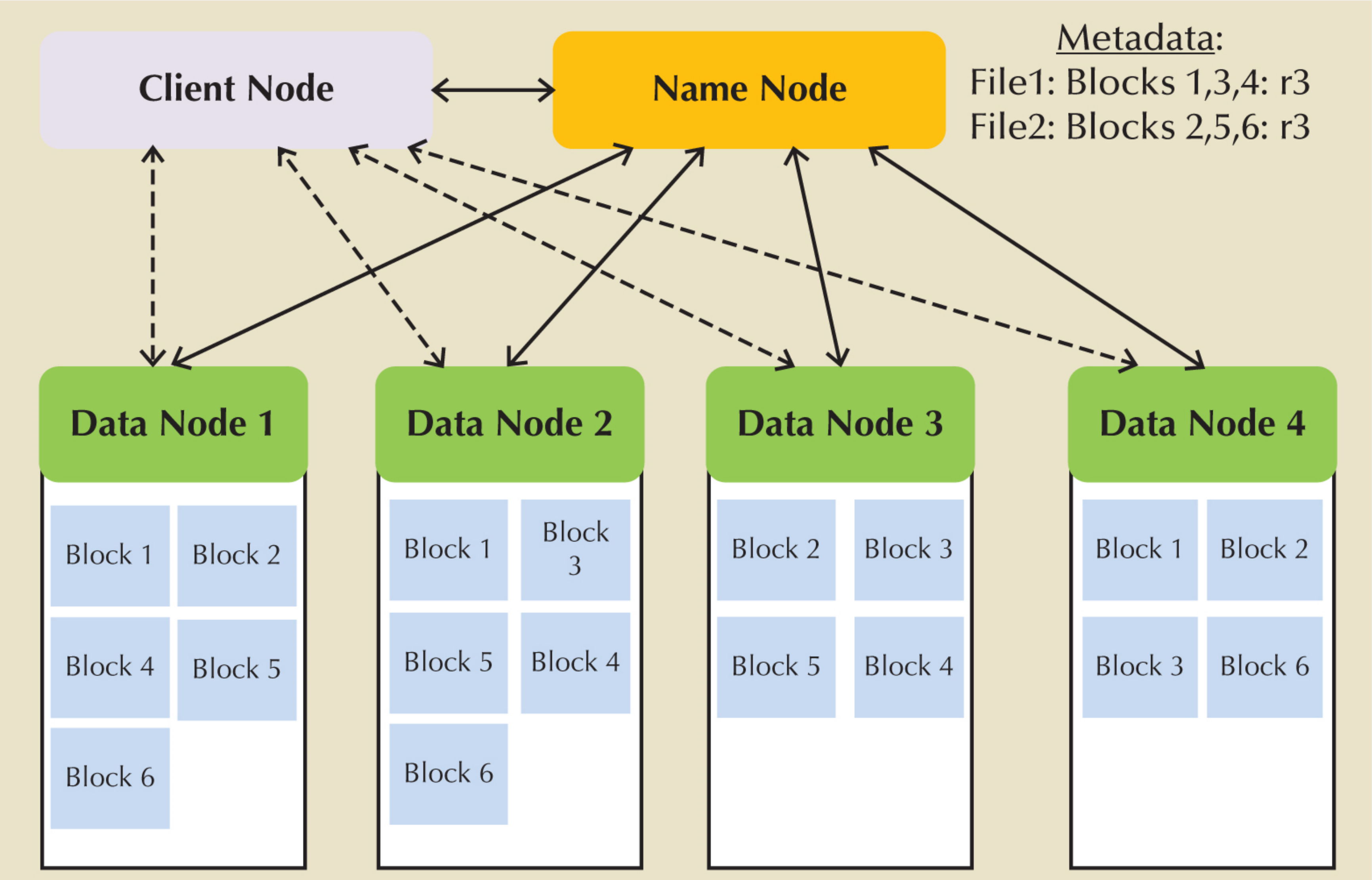
Hadoop - HDFS

- Fault tolerance
  - *data replication; so, when one device fails, data is available on another device*

Hadoop – Nodes (computers)

- Data node stores the actual file data
- Name node contains file system metadata
- Client node makes requests to the file system
- Data node communicates with name node by regularly sending block reports and heartbeats

Hadoop – Nodes (computers)



Hadoop – MapReduce

- A complex task is split into smaller subtasks, perform the subtasks, and produce a final result
- Map takes a collection of data and sorts and filters it into a set of **key-value pairs**
- Reduce **combines results of map function** to produce a **single result**

NoSQL

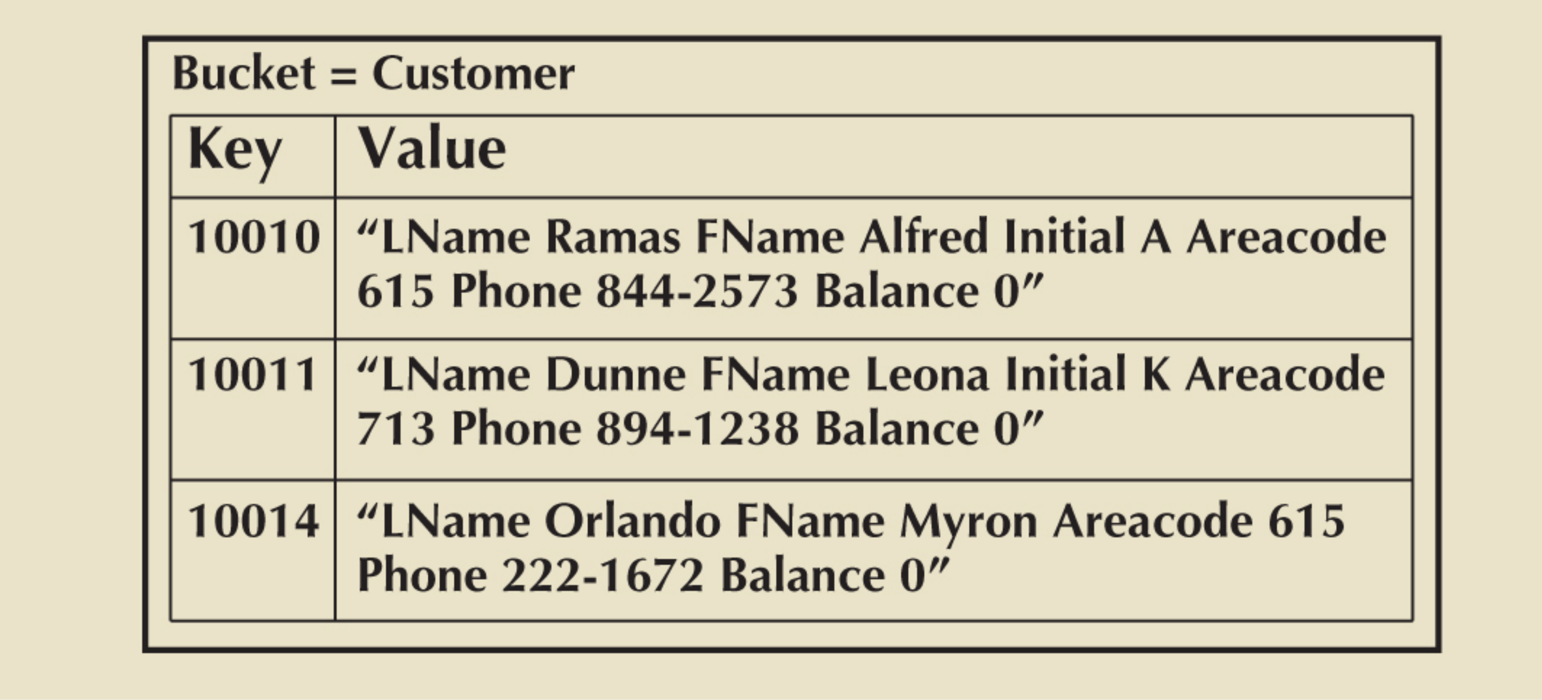
NoSQL is just a name of non-relational DB technologies that address Big Data

1. Key-value Databases
2. Document Databases
3. Column-oriented Databases
4. Graph Databases

https://www.youtube.com/watch?v=BgQFJ_UNIgw

NoSQL - Key-Value Databases

- Data is stored as a collection of key-value pairs

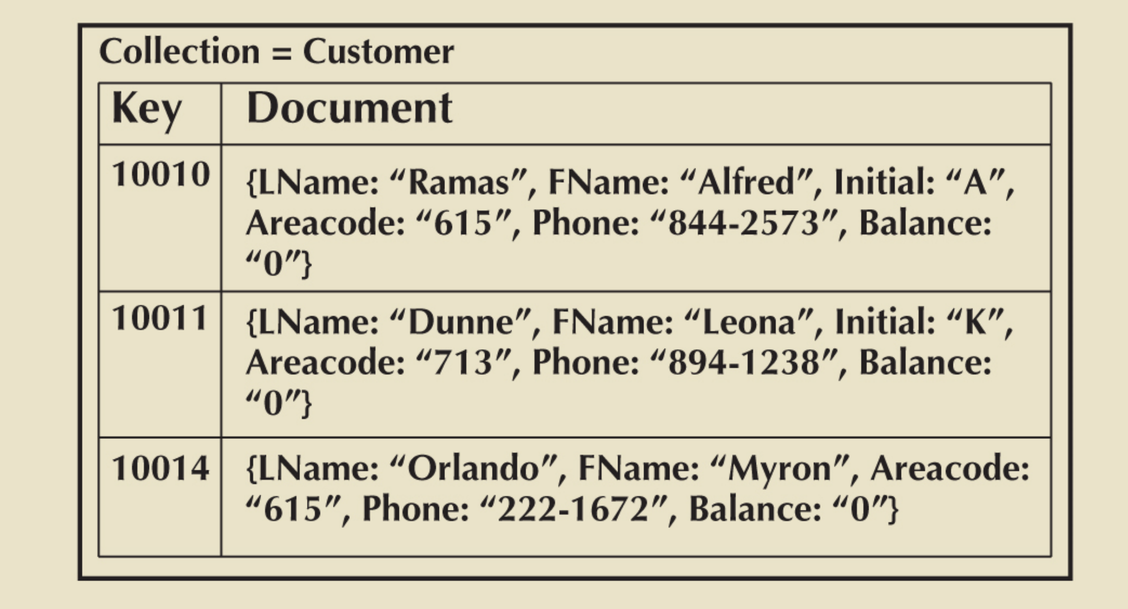


Bucket = Customer

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"

NoSQL - Document Databases

- Store data in key-value pairs
- Value components are tag-encoded documents
- Find all documents based on a tag/value

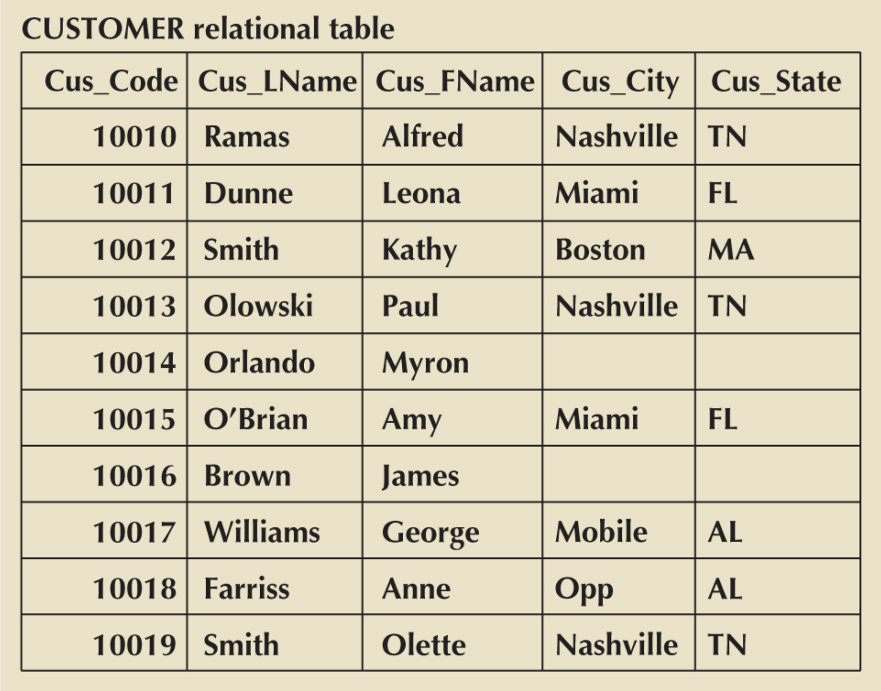


Collection = Customer

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”}

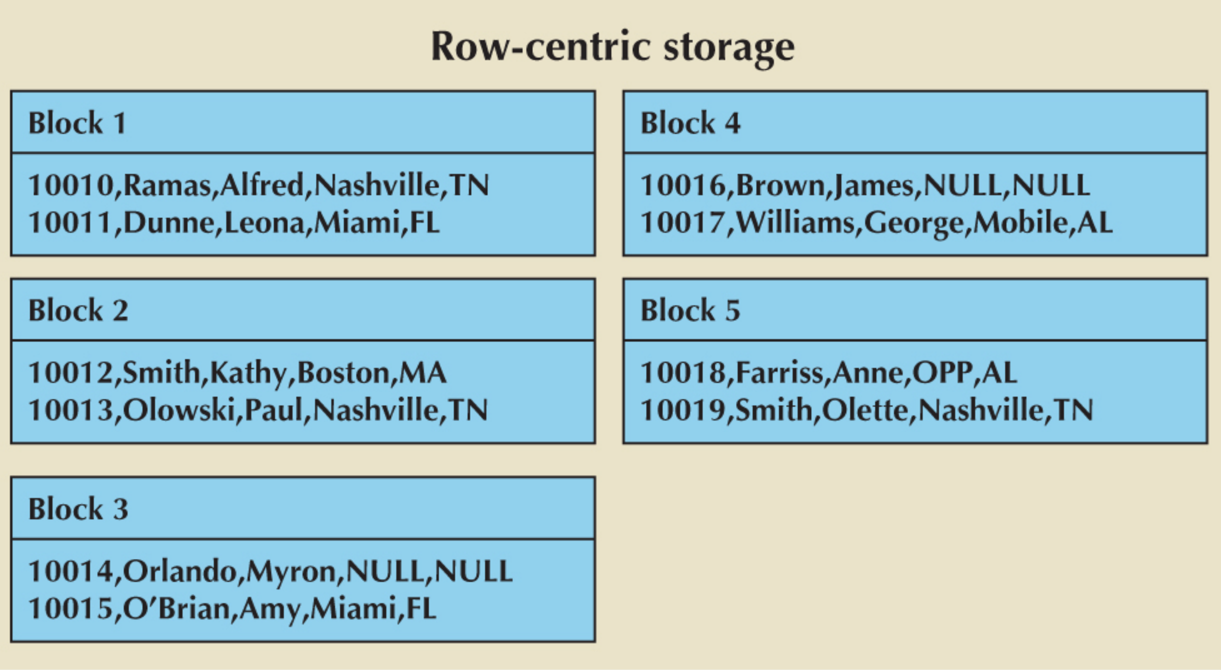
NoSQL - Column-Oriented DBs

- Column-Centric Storage
  - data stored in blocks which hold data from a single column across many 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

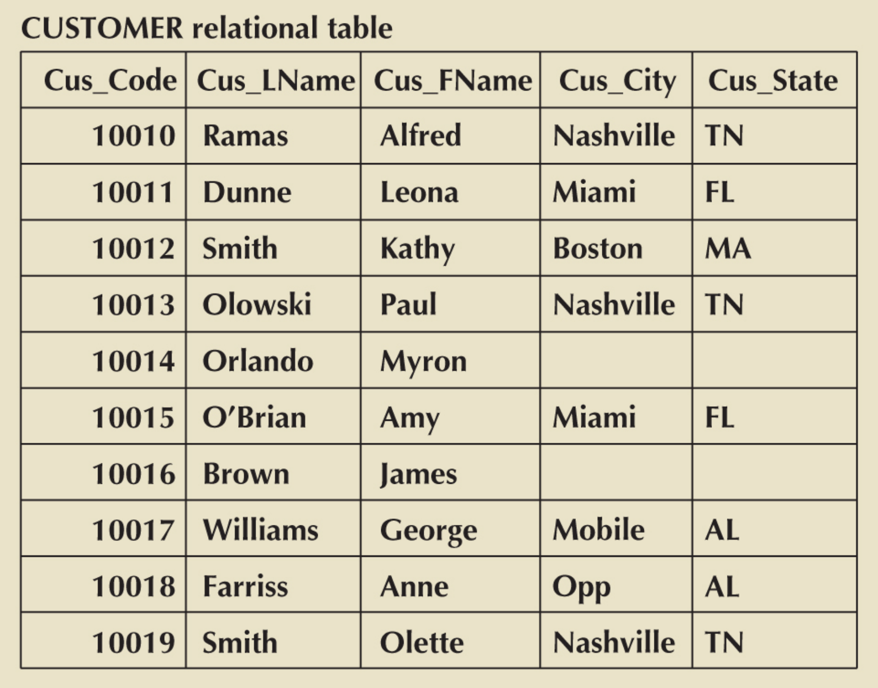
10016,Brown,James,NULL,NULL
10017,Williams,George,Mobile,AL

Block 5

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

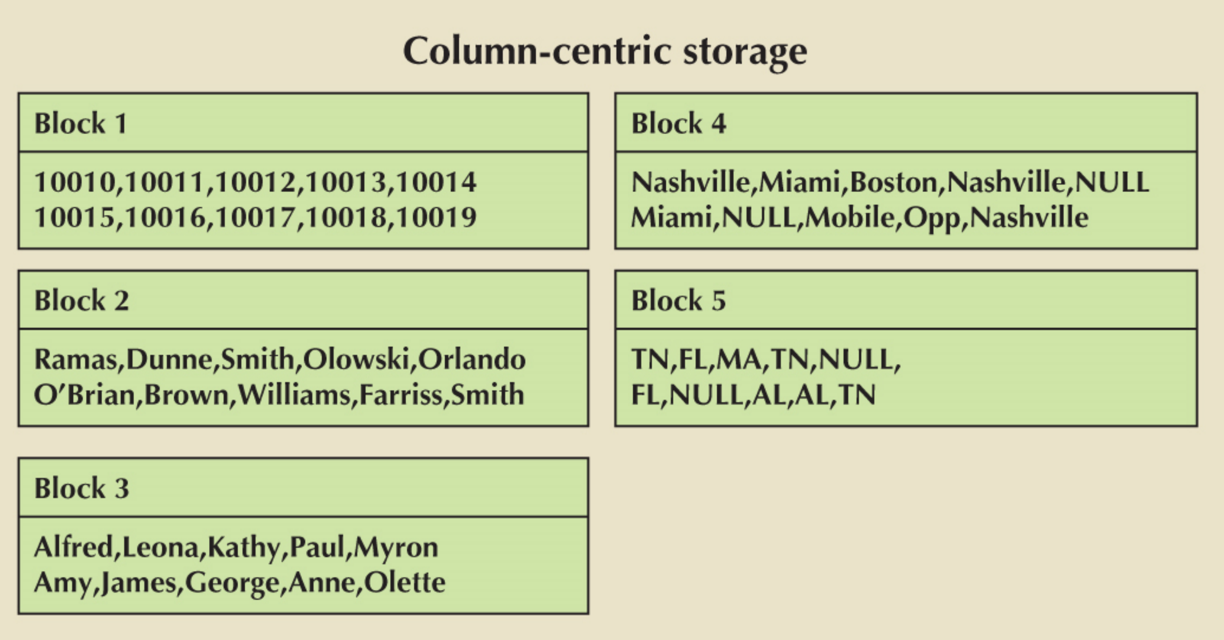
NoSQL - Column-Oriented DBs

- Row-Centric Storage
  - data stored in blocks 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



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

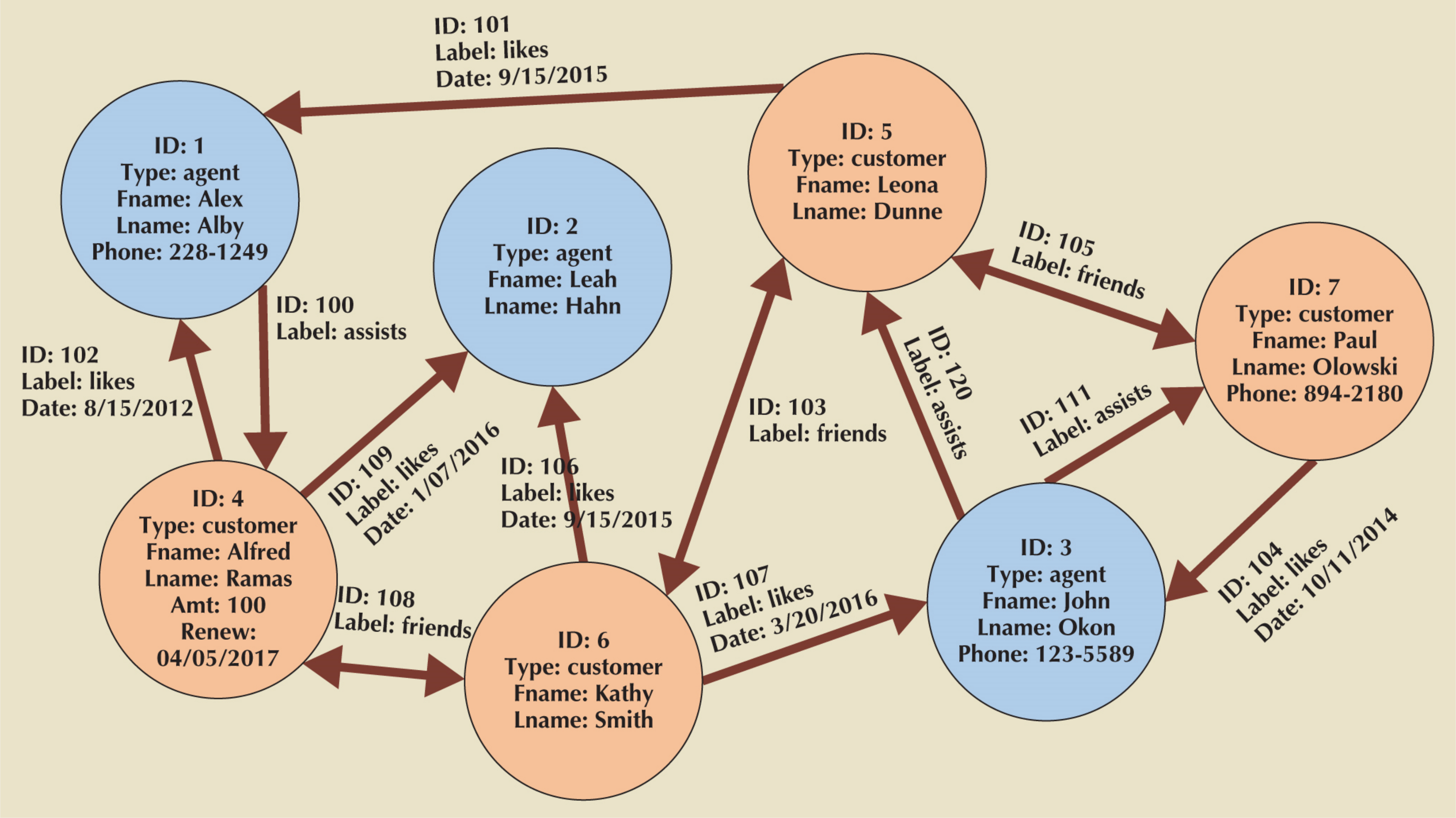
Block 5

TN,FL,MA,TN,NULL,
FL,NULL,AL,AL,TN

NoSQL - Graph Databases

- Data is stored on relationship-rich data as a collection of nodes and edges
  - Nodes and edges have attributes/values
  - Traversal is a query in a graph database

NoSQL - Graph Databases



ID: 101
Label: likes
Date: 9/15/2015

ID: 1
Type: agent
Fname: Alex
Lname: Alby
Phone: 228-1249

ID: 2
Type: agent
Fname: Leah
Lname: Hahn

ID: 5
Type: customer
Fname: Leona
Lname: Dunne

ID: 105
Label: friends

ID: 7
Type: customer
Fname: Paul
Lname: Olowski
Phone: 894-2180

ID: 100
Label: assists

ID: 102
Label: likes
Date: 8/15/2012

ID: 120
Label: assists

ID: 103
Label: friends

ID: 111
Label: assists

ID: 109
Label: likes
Date: 1/07/2016

ID: 106
Label: likes
Date: 9/15/2015

ID: 4
Type: customer
Fname: Alfred
Lname: Ramas
Amt: 100
Renew:
04/05/2017

ID: 108
Label: friends

ID: 6
Type: customer
Fname: Kathy
Lname: Smith

ID: 107
Label: likes
Date: 3/20/2016

ID: 3
Type: agent
Fname: John
Lname: Okon
Phone: 123-5589

ID: 104
Label: likes
Date: 10/11/2014

Summary

1. Dig Data (V, V, V, …)
2. Hadoop
   - *Hadoop Distributed File System*
   - *MapReduce*
3. NoSQL Databases
   - *Key-Value Databases*
   - *Document Databases*
   - *Column-Oriented DBs*
   - *Graph Databases*