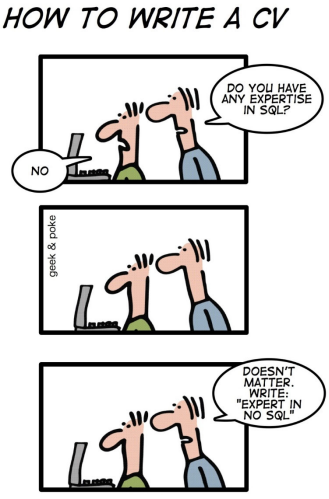
XIV. Big Data and NoSQL

NoSQL Databases

- Recall the three V's of "big data"
  - Volume
  - Velocity
  - Variety



Big Data

- Volume: quantity of data to be stored
  - Scaling up: keeping the same number of systems but migrating each one to a larger system
  - Scaling out: when the workload exceeds server capacity, it is spread out across a number of servers
- Velocity: speed at which data is entered into system and must be processed
  - Stream processing: focuses on input processing and requires analysis of data stream as it enters the system
  - Feedback loop processing: analysis of data to produce actionable results
- Variety: variations in the structure of data to be stored
  - Structured data: fits into a predefined data model
  - Unstructured data: does not fit into a predefined model

Big Data

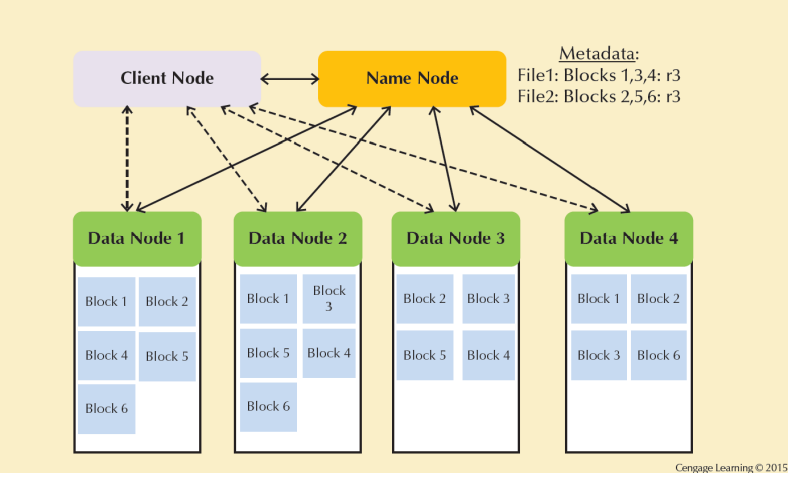
- Other characteristics
  - Variability: changes in meaning of data based on context
    - Sentiment analysis: attempts to determine if a statement conveys a positive, negative, or neutral attitude about a topic
  - Veracity: trustworthiness of data
  - Value: degree data can be analyzed for meaningful insight
  - Visualization: ability to graphically resent data to make it understandable
- Relational databases are not necessarily the best for storing and managing all organizational data
  - Polyglot persistence: coexistence of a variety of data storage and management technologies within an organization's infrastructure

Key Assumptions of Hadoop Distributed File System

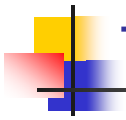
- High volume
- Write-once, read-many
- Streaming access
- Move computations to the data
- Fault tolerance

5

Hadoop Distributed File System (HDFS)



6



Types of NoSQL Databases

- NoSQL: non-relational database technologies developed to address Big Data challenges
  - Name does not describe what the NoSQL technologies are, but rather what they are not (poor job of that as well)
- Key-value (KV) databases: conceptually the simplest of the NoSQL data models
  - Store data as a collection of key-value pairs organized as buckets which are the equivalent of tables
- Document databases: similar to key-value databases and can almost be considered a subtype of KV databases
  - Store data in key-value pairs in which the value components are encoded documents grouped into large groups called collections

Types of NoSQL Databases

FIGURE 14.7 KEY-VALUE DATABASE STORAGE

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"

Types of NoSQL Databases

FIGURE 14.8 DOCUMENT DATABASE TAGGED FORMAT

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"}

Types of NoSQL Databases

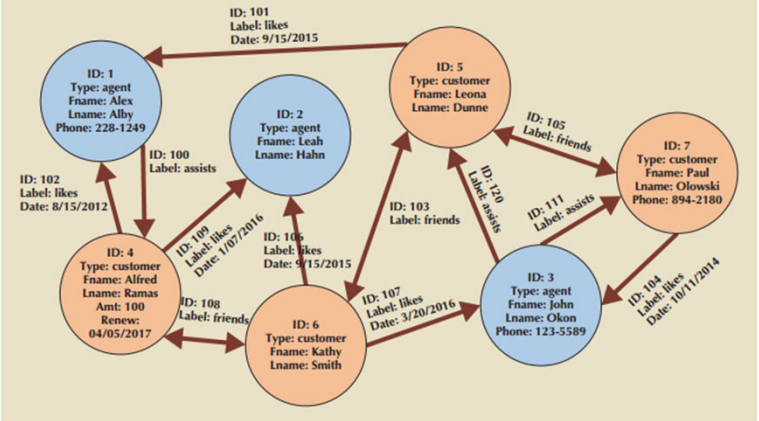
- Column family databases
  - Organizes data in key-value pairs with keys mapped to a set of columns in the value component
  - Columns can loosely be thought of as attributes from the relational model
- Graph databases store data on relationship-rich data as a collection of nodes and edges
  - Store data about relationship-rich environments
  - Nodes, edges, and properties
    - Properties: like attributes; they are the data that we need to store about the node or edge
  - Traversal: query in a graph database

Types of NoSQL Databases

Column Family Name	CUSTOMERS
Key	Rowkey 1
Columns	City: Nashville Fname: Alfred Lname: Ramas State: TN
Key	Rowkey 2
Columns	Balance: 345.86 Fname: Kathy Lname: Smith
Key	Rowkey 3
Columns	Company: Local Markets, Inc. Lname: Dunne

Types of NoSQL Databases

FIGURE 14.11 GRAPH DATABASE REPRESENTATION

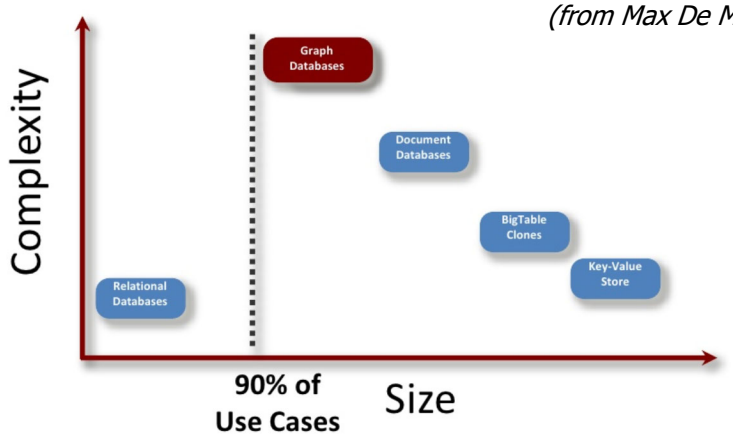


Types of NoSQL Databases

- Aggregate awareness: data is collected or aggregated around a central topic or entity
  - Aggregate aware database models achieve clustering efficiency by making each piece of data relatively independent
- Graph databases, like relational databases, are aggregate ignorant
  - Do not organize the data into collections based on a central entity

Types of NoSQL Databases

(from Max De Marzi)



14

Rank			DBMS	Database Model
Dec 2019	Nov 2019	Dec 2018		
1.	1.	1.	Oracle	Relational, Multi-model
2.	2.	2.	MySQL	Relational, Multi-model
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model
4.	4.	4.	PostgreSQL	Relational, Multi-model
5.	5.	5.	MongoDB	Document, Multi-model
6.	6.	6.	IBM Db2	Relational, Multi-model
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model
8.	8.	↓ 7.	Redis	Key-value, Multi-model
9.	9.	9.	Microsoft Access	Relational
10.	10.	↑ 11.	Cassandra	Wide column
11.	11.	↓ 10.	SQLite	Relational
12.	12.	12.	Splunk	Search engine
13.	13.	↑ 14.	MariaDB	Relational, Multi-model
14.	14.	↑ 15.	Hive	Relational
15.	15.	↓ 13.	Teradata	Relational, Multi-model
16.	16.	↑ 21.	Amazon DynamoDB	Multi-model
17.	17.	↓ 16.	Solr	Search engine

NewSQL Databases

- Database model that attempts to provide ACID-compliant transactions across a highly distributed infrastructure
  - Latest technologies to appear in the data management area to address Big Data problems
  - No proven track record
  - Have been adopted by relatively few organizations

NewSQL Databases

- NewSQL databases support:
  - SQL as the primary interface
  - ACID-compliant transactions
- Similar to NoSQL, NewSQL databases also support:
  - Highly distributed clusters
  - Key-value or column-oriented data stores

MongoDB Example

- Popular document database
  - Among the NoSQL databases currently available, MongoDB has been one of the most successful in penetrating the database market
- MongoDB, comes from the word humongous as its developers intended their new product to support extremely large data sets
  - High availability
  - High scalability
  - High performance

MongoDB Example

- Importing Documents in MongoDB
  - Import JSON, CSV, XML, and other file types
- Example of a MongoDB Query Using find()
  - Methods are programmed functions to manipulate objects
    - Find() method retrieves objects from a collection that match the restrictions provided
    - Pretty() method is used to improve readability of the documents by placing key:value pairs on separate lines

MongoDB Example

20

Summary

- Big Data is characterized by data of such volume, velocity, and/or variety that the relational model struggles to adapt to it
- Volume, velocity, and variety are collectively referred to as the 3 Vs of Big Data
- The Hadoop framework has quickly emerged as a standard for the physical storage of Big Data
- NoSQL is a broad term to refer to any of several nonrelational database approaches to data management
- Key-value databases store data in key-value pairs
- Document databases also store data in key-value pairs, but the data in the value component is an encoded document

Summary

- Column-oriented databases, also called column family databases, organize data into key-value pairs in which the value component is composed of a series of columns, which are themselves key-value pairs
- Graph databases are based on graph theory and represent data through nodes, edges, and properties
- NewSQL databases attempt to integrate features of both RDBMS (providing ACID-compliant transactions) and NoSQL databases (using a highly distributed infrastructure)
- MongoDB is a document database that stores documents in JSON format
- Neo4j is a graph database that stores data as nodes and relationships, both of which can contain properties to describe them