## **Table of Contents**

	Storage and Retrieval	
Pref	Data Structures That Power Your Database	xiii
	SS Labites and LSM-Trees	17
Par	t I. Foundations of Data Systems	171
1.	Reliable, Scalable, and Maintainable Applications	3
	Thinking About Data Systems	4
	Reliability Reliability Reliability	6
	Hardware Faults	7
	Software Errors	8
	Human Errors	9
	How Important Is Reliability?	10
	Scalability Scalability Scalability	10
	Describing Load	
	Describing Performance	13
	Approaches for Coping with Load	
	Maintainability Amu and ambon E to alamon	
	Operability: Making Life Easy for Operations	19
	Simplicity: Managing Complexity	20
	Evolvability: Making Change Easy	21
	Summary	22
2.	Data Models and Query Languages	. 27
	Relational Model Versus Document Model	28
	The Birth of NoSQL	29
	The Object-Relational Mismatch	29
	Many-to-One and Many-to-Many Relationships	33
	Are Document Databases Repeating History?	36

	Relational Versus Document Databases Today	38
	Query Languages for Data	42
	Declarative Queries on the Web	44
	MapReduce Querying	46
	Graph-Like Data Models	49
	Property Graphs	50
	The Cypher Query Language	52
	Graph Queries in SQL	53
	Triple-Stores and SPARQL	55
	The Foundation: Datalog	60
	Summary	63
3.	Storage and Retrieval	. 69
	Data Structures That Power Your Database	70
	Hash Indexes	72
	SSTables and LSM-Trees	76
	B-Trees	79
	Comparing B-Trees and LSM-Trees Trees and LSM-Trees	83
	Other Indexing Structures	85
	Transaction Processing or Analytics? Will Manuscript Ma	90
	Data Warehousing	91
	Stars and Snowflakes: Schemas for Analytics	93
	Column-Oriented Storage	95
	Column Compression	97
	Sort Order in Column Storage	99
	Writing to Column-Oriented Storage	101
	Aggregation: Data Cubes and Materialized Views	101
	Summary bad I gridings Cl	103
4.	Encoding and Evolution	111
	Formats for Encoding Data villdsatsmisM	112
	Language-Specific Formats	113
	JSON, XML, and Binary Variants	114
	Thrift and Protocol Buffers was a synd of protocol graduated and protocol Buffers	117
	Summary	122
	The Merits of Schemas	127
	Modes of Dataflow	128
	Dataflow Through Databases IsboM inserusoff aners V IsboM Innoins ST	129
	Dataflow Through Services: REST and RPC	131
	Message-Passing Dataflow	136
	Many-to-One and Many-to-Many Relationships	139

-			PERDAN	
Part II.	Distributed Data			
Bank Com				

Leaders and Followers Synchronous Versus Asynchronous Replication 153 Setting Up New Followers Handling Node Outages Implementation of Replication Logs Problems with Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads 152 153 154 155 156 167 168 169 169 169 160 160 160 161 162 165	5	Replication	151
Synchronous Versus Asynchronous Replication Setting Up New Followers Handling Node Outages Implementation of Replication Logs Implementation of Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Servendary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary 216 7. Transactions. 221	٥.		152
Setting Up New Followers Handling Node Outages Implementation of Replication Logs Problems with Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning, Partitioning and Replication Partitioning by Key Range Partitioning by Key Range Partitioning Secondary Indexes Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary 216 7. Transactions. 221 217 218 218 218 219 219 211 210 210 210 210 210 210 210 210 210		D 1:	153
Handling Node Outages Implementation of Replication Logs Problems with Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Hulti-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning Partitioning and Replication Partitioning of Key-Value Data Partitioning of Key-Value Data Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  216 7. Transactions.  217  168  168  169  169  169  169  167  168  168  169  169  169  169  169  169			155
Implementation of Replication Logs Problems with Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning Partitioning and Replication Partitioning of Key-Value Data Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  7. Transactions.  163 165 165 165 167 168 168 167 168 168 169 169 169 169 169 169 160 168 169 169 169 169 169 169 169 169 169 169		Handling Node Outages	156
Problems with Replication Lag Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  7. Transactions.  162 162 163 164 165 165 166 167 168 168 168 169 169 169 169 169 169 169 169 169 169		Implementation of Replication Logs	
Reading Your Own Writes Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. 199 Partitioning and Replication Partitioning by Key Range Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  160  168 167 168 168 169 168 168 169 168 168 169 168 168 169 168 168 168 168 168 168 168 168 168 168			161
Monotonic Reads Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  7. Transactions.  211  168 167 167 168 168 168 169 168 169 169 169 169 169 175 175 175 175 175 175 175 175 175 175		2000년 1일 전 1000년 1일 전 1000년 1일	162
Consistent Prefix Reads Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning by Key-Value Data Partitioning by Key-Value Data Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  168 168 168 168 168 168 168 168 168 16			164
Solutions for Replication Lag Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Hulti-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  168 168 168 168 168 168 168 168 168 16			165
Multi-Leader Replication Use Cases for Multi-Leader Replication Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  168 168 168 168 168 168 168 168 168 16		[2] 10 - 10 - 10 - 10 - 10 - 10 - 10 - 10	167
Use Cases for Multi-Leader Replication Handling Write Conflicts Hulti-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning by Hash of Key Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  168 171 177 177 177 177 177 177 177 177 17			1/0
Handling Write Conflicts Multi-Leader Replication Topologies Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning Partitioning and Replication Partitioning of Key-Value Data Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  177 177 177 177 177 177 177 177 177 1			
Multi-Leader Replication Topologies  Leaderless Replication  Writing to the Database When a Node Is Down  Limitations of Quorum Consistency  Sloppy Quorums and Hinted Handoff  Detecting Concurrent Writes  Summary  192  6. Partitioning.  Partitioning and Replication  Partitioning of Key-Value Data  Partitioning by Key Range  Partitioning by Hash of Key  Skewed Workloads and Relieving Hot Spots  Partitioning Secondary Indexes  Partitioning Secondary Indexes by Document  Partitioning Secondary Indexes by Term  Rebalancing Partitions  Strategies for Rebalancing  Operations: Automatic or Manual Rebalancing  Request Routing  Parallel Query Execution  Summary  221  7. Transactions.  231  241  242  241  242  243  244  246  247  248  248  249  240  240  240  241  240  241  241  241			171
Leaderless Replication Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary 192  6. Partitioning. 199 Partitioning and Replication Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary 221  7. Transactions. 281 281 282 283 284 285 286 287 288 288 288 288 288 288 288 288 288			
Writing to the Database When a Node Is Down Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary 192  6. Partitioning			177
Limitations of Quorum Consistency Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes 184 Summary 192 6. Partitioning. Partitioning and Replication Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary 216 7. Transactions. 218 218 28 290 201 202 203 204 205 205 205 205 205 205 206 207 208 208 209 210 210 210 211 210 211 211 211 211 211			177
Sloppy Quorums and Hinted Handoff Detecting Concurrent Writes Summary  6. Partitioning			181
Detecting Concurrent Writes Summary  6. Partitioning			183
Summary  6. Partitioning			
Partitioning and Replication Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  200 201 202 203 204 205 206 206 207 208 208 208 209 209 209 209 209 209 209 210 209 211 210 210 211 211 211 211 211 211 211			192
Partitioning and Replication Partitioning of Key-Value Data  Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  200 201 202 203 204 205 206 206 207 208 208 208 209 209 209 209 209 209 209 209 210 209 210 210 210 211 210 211 211 211 211 211	100		100
Partitioning of Key-Value Data Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Partitioning Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  201 202 203 204 205 206 206 207 208 208 209 209 209 209 209 210 209 211 210 211 211 211 212 213 214 216 216 216 216 217 217 218 221	6.	30_14 전문 전 20 CHONG CONTROL CO	200
Partitioning by Key Range Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  202 203 205 206 206 207 208 208 209 209 210 209 211 209 212 213 214 214 216 216 216 216 216 217 217 218 221			201
Partitioning by Hash of Key Skewed Workloads and Relieving Hot Spots Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  203 205 206 207 208 208 208 209 209 210 210 210 211 211 211 212 213 214 216 216 216 216 217 217 218 221			
Skewed Workloads and Relieving Hot Spots  Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  205 206 Partitioning Secondary Indexes by Term 208 Rebalancing Partitions 209 210 211 212 213 214 216 216 216 216 217 217 218 221		## # [2] - 12 - 12 - 12 - 13 - 13 - 13 - 13 - 13	
Partitioning and Secondary Indexes Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Parallel Query Execution Summary  7. Transactions.		Cl 1747 11 1 1 1 Delining Heat	
Partitioning Secondary Indexes by Document Partitioning Secondary Indexes by Term Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Request Routing Parallel Query Execution Summary  7. Transactions.			
Partitioning Secondary Indexes by Term  Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Request Routing Parallel Query Execution Summary  7. Transactions.			201
Rebalancing Partitions Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Request Routing Parallel Query Execution Summary  7. Transactions.			200
Strategies for Rebalancing Operations: Automatic or Manual Rebalancing Request Routing Parallel Query Execution Summary  7. Transactions.			200
Operations: Automatic or Manual Rebalancing Request Routing Parallel Query Execution Summary  7. Transactions.		0	
Request Routing Parallel Query Execution Summary  7. Transactions.		10.1.1	
Parallel Query Execution 216 Summary 216 7. Transactions. 221			
7. Transactions. 216			
7. Transactions			210
7. Transactions			210
	7		221
	-		

	The Meaning of ACID		223
	Single-Object and Multi-Object Operations		228
	Weak Isolation Levels		233
	Read Committed		234
	Snapshot Isolation and Repeatable Read		237
	Preventing Lost Updates		242
	Write Skew and Phantoms		246
	Serializability		251
	rictan beria Execution		252
	Two-Phase Locking (2PL)		257
	Serializable Snapshot Isolation (SSI)		261
	Summary		266
8.	The Trouble with Distributed Systems	a.l. goll salgast tol anotolos.	. 273
	Faults and Partial Failures		274
	Cloud Computing and Supercomputing		275
	Unreliable Networks	Handling Write Conflicts	277
	Network Faults in Practice	Multi-Leader Replication To	279
	Detecting Faults		280
	Timeouts and Unbounded Delays		281
	Synchronous Versus Asynchronous Network	ks Immove to anothermid	284
	Unreliable Clocks		287
	Monotonic Versus Time-of-Day Clocks		288
	Clock Synchronization and Accuracy		289
	Relying on Synchronized Clocks		291
	Process Pauses		295
	Knowledge, Truth, and Lies		300
	The Truth Is Defined by the Majority		300
	Byzantine Faults		304
	System Model and Reality		306
	Odililial		310
0			321
	consistency and consensus	Partitioning Secondary Inde	322
	Linearizability		
	What Makes a System Linearizable?		325
	Relying on Linearizability		330
	Implementing Linearizable Systems		
	The Cost of Linearizability		335
	Ordering Guarantees		
	Ordering and Causality		339
	Sequence Number Ordering		343
	orquence runner ordering		515

Total Order Broadcast	348
10	352
Atomic Commit and Two-Phase Commit (2PC)	354
Distributed Transactions in Practice	360
Fault-Tolerant Consensus	211
Membership and Coordination Services	0.00
Summary wollsta' bnuor A enottabiliga A gai	272
ing Derived State	
or Correctness	
Part III. Derived Data	
10. Batch Processing	389
Batch Processing with Unix Tools	391
Simple Log Analysis	391
The Unix Philosophy	394
MapReduce and Distributed Filesystems	
MapReduce Job Execution	399
Reduce-Side Joins and Grouping	403
Map-Side Joins	408
The Output of Batch Workflows	411
Comparing Hadoop to Distributed Databases	414
Beyond MapReduce	419
Materialization of Intermediate State	419
Graphs and Iterative Processing	424
High-Level APIs and Languages	120
Summary	429
A Journal of the need to be again that appointed the only and respond	
11. Stream Processing	439
Transmitting Event Streams	440
Messaging Systems	441
Partitioned Logs	446
Databases and Streams	451
Keening Systems in Sync	452
Change Data Capture	454
Event Sourcing	457
State, Streams, and Immutability	459
Processing Streams	464
Uses of Stream Processing	465
Reasoning About Time	468
Stream Joins	472
Fault Tolerance	476
Summary	479

12.	The Future of Data Systems	Peobeorii, rebrΩ letoT.	. 489	
	Data Integration		490	
	Combining Specialized Tools by Deriving Data			
	Batch and Stream Processing	Distributed Transactions in Practice	494	
	Unbundling Databases		499	
	Composing Data Storage Technolog	Membership and Coordination	499	
	Designing Applications Around Dataflow			
	Observing Derived State		504 509	
	Aiming for Correctness		515	
	The End-to-End Argument for Datal Enforcing Constraints	bases synd beying	516 521	
	Timeliness and Integrity		524	
	Trust, but Verify			
	Doing the Right Thing		533	
	Predictive Analytics		533	
	Privacy and Tracking			
	Summary	MapReduce Job Execution	543	
	Network Faults to Practice	Reduce Side Joins and Grouping	343	
Glos	sary		. 553	
LIE	Limeaus and Unbounded Delays 45	The Output of Batch Workflows	281	
Inde	х	Comparing Hadoon to Distributed:	. 559	
611	Intellable Clocks	eyond MapReduce		
		tream Processing		