Databases and information management

Lecture 10

Piotr Syga

2022-12-08

Indexes

- similarly to SQL databases, Mongo supports indexes
- indexes are created at the collection level (cf. table's indexes)
- there's always at least 1 index – _id
- index creation `db.collection.createIndex( <key and index type specification>, <options> )`
- possible option include `unique`, `name`, `sparse`, `expireAfterSeconds`
- the index may be created on a single or multiple fields, the default name is equal to concatenation of all the used fields
- one can determine the order of the index by putting a positive or a negative key, e.g., `db.students.createIndex({surname:-1})`
- the order of fields in multifield indexes is important, moreover a separate index is created for each level of indexed embedded documents or arrays
- partial vs sparse index

Indexes

Limitation:

- there can be only 1 index of a given name
- if the unique option is set to true, the index will prohibit to enter duplicate values
- a single collection is limited to 64 indexes (including those for embedded documents and arrays)
- there can be no more than 32 fields in a compound index
- the limit on the memory usage of `CreateIndexes` is 200MB, overhead has to be done in an additional command

Transactions

- MongoDB uses atomic operations on a single document, transactions are required for atomicity of multidocument queries
- transactions may be used on replicas and sharded sets
- one can create collections inside a transactions (both implicitly as a side effect of evocation and explicitly)
- indexes may be created inside a transactions but only on non-existent or empty collections
- race condition may require to use write concern majority
- the transaction is started by creating a session object `mySession = db.getMongo().startSession()` and then starting transaction `mySession.startTransaction()`
- the transaction is ended with `mySession.commitTransaction()`

Replication

- replicas are a copies of the data kept on a separate process (and possibly server)
- the main goal is to provide redundancy (for majority votes and backups) and high availability
- replicas consist of a primary database copy and a set (at least two) secondary copies that allow data restoration – the votes are based on majority for inconsistent databases
- in the case of lacking replicas one could establish arbiter who cannot become a primary, however takes part in all the votes

Creating a replica set

We create the set in command line `mongod -port <port> -dbpath /var/lib/mongodb -replSet <name>`.
Starting the set and adding new instances are performed in mongo shell `rs.initiate()`, `rs.add(<hostname:port>)`

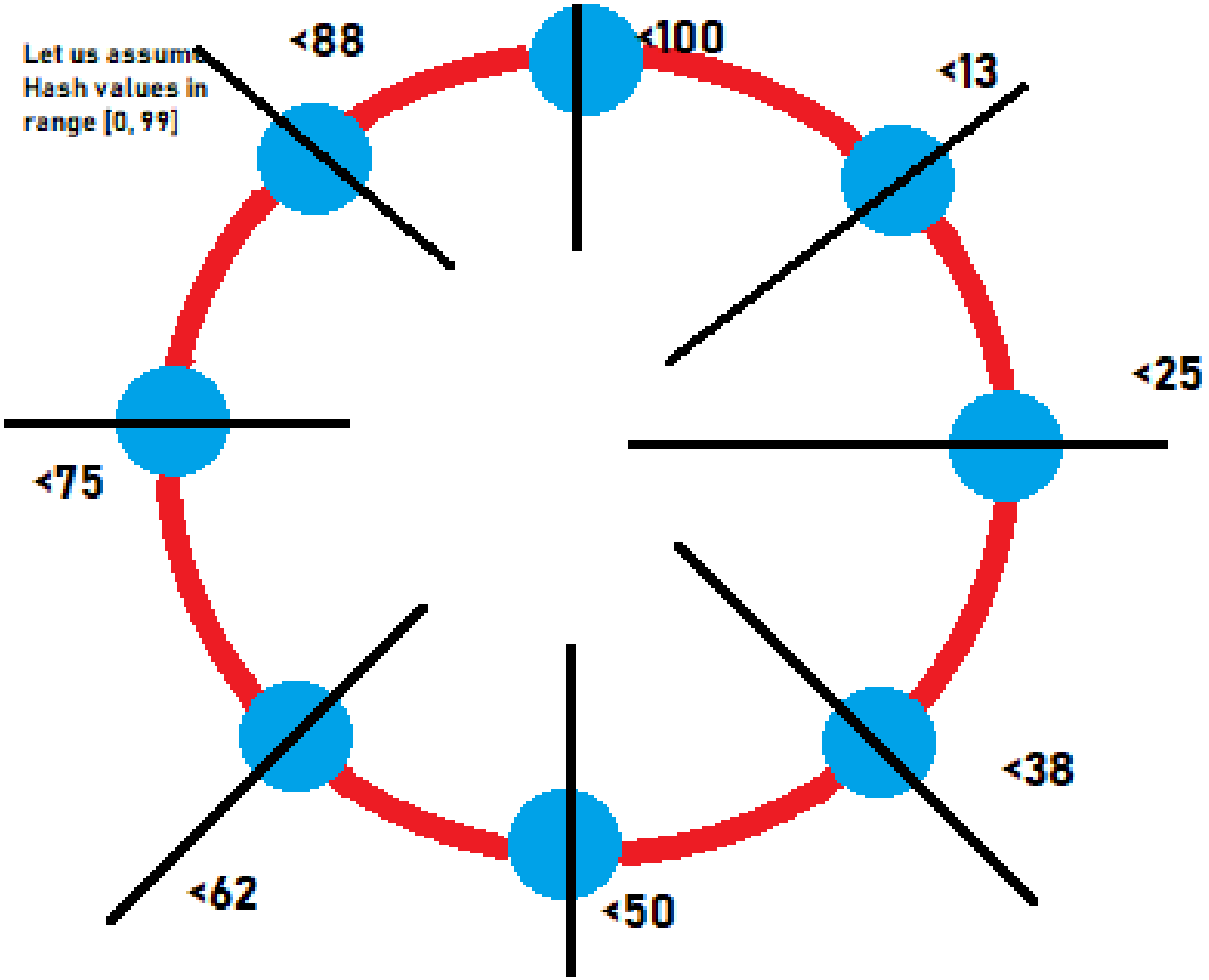
Main properties

- NOSQL columnstore database
- runs on Java Virtual Machine and requires access to a local disk
- each row in a column is stored as a key–value pair
- designed for simplicity and horizontal scaling
- peer-to-peer architecture
- user defined limits of availability and consistency
- fast writes and reads
- supports transactions
- based on tables with flexible schema
- CQL similar to SQL dialects
- *query first* approach
- allows no JOINs
- utilizes Bloom filters(with specified false positive probability), column and table compression and dynamic gossiping

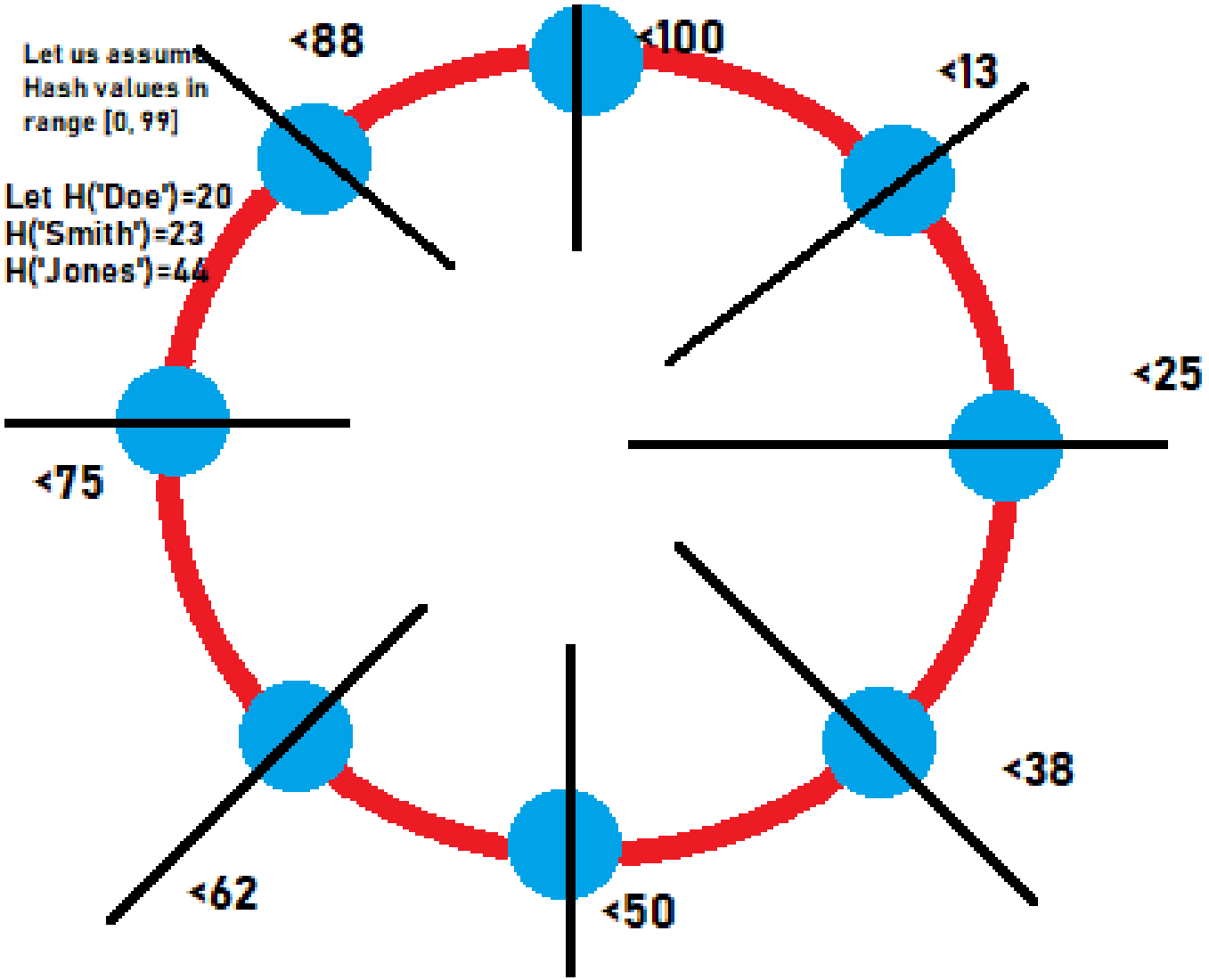
Ring Architecture

- a single keyspace ('equivalent' to SQL database) may be kept on multiple nodes
- data distribution is performed with respect to table's *partition key*
- 64bit Hash function assigns data specific cluster or a node
- each node is assigned token range up to some threshold, in an ordered way in a modulo ring
- each range may be divided into virtual nodes (multiple sections inside one base range)
- the ranges and the number of virtual nodes do not have to be equal
- by default all nodes have 256 virtual nodes

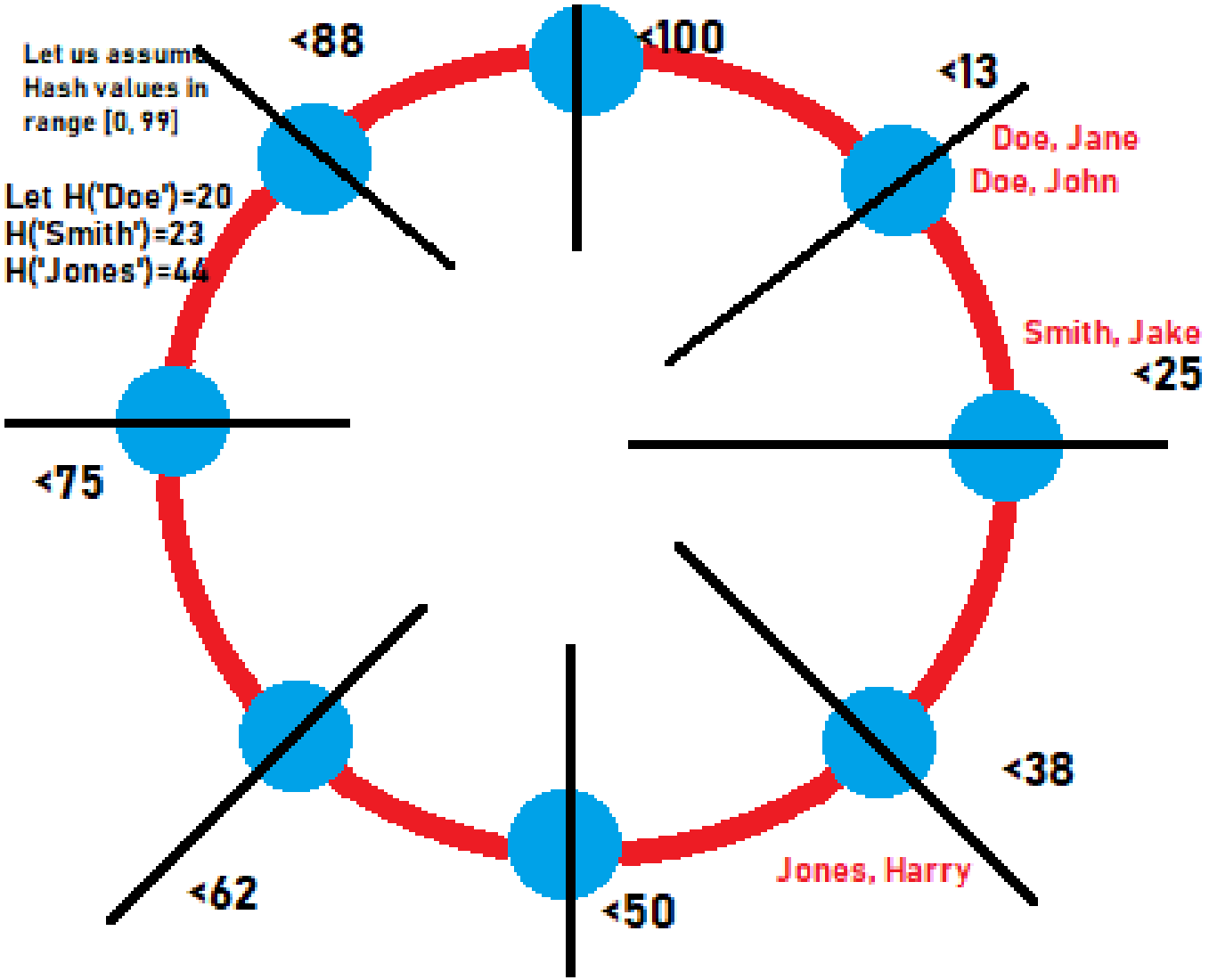
Ring Architecture



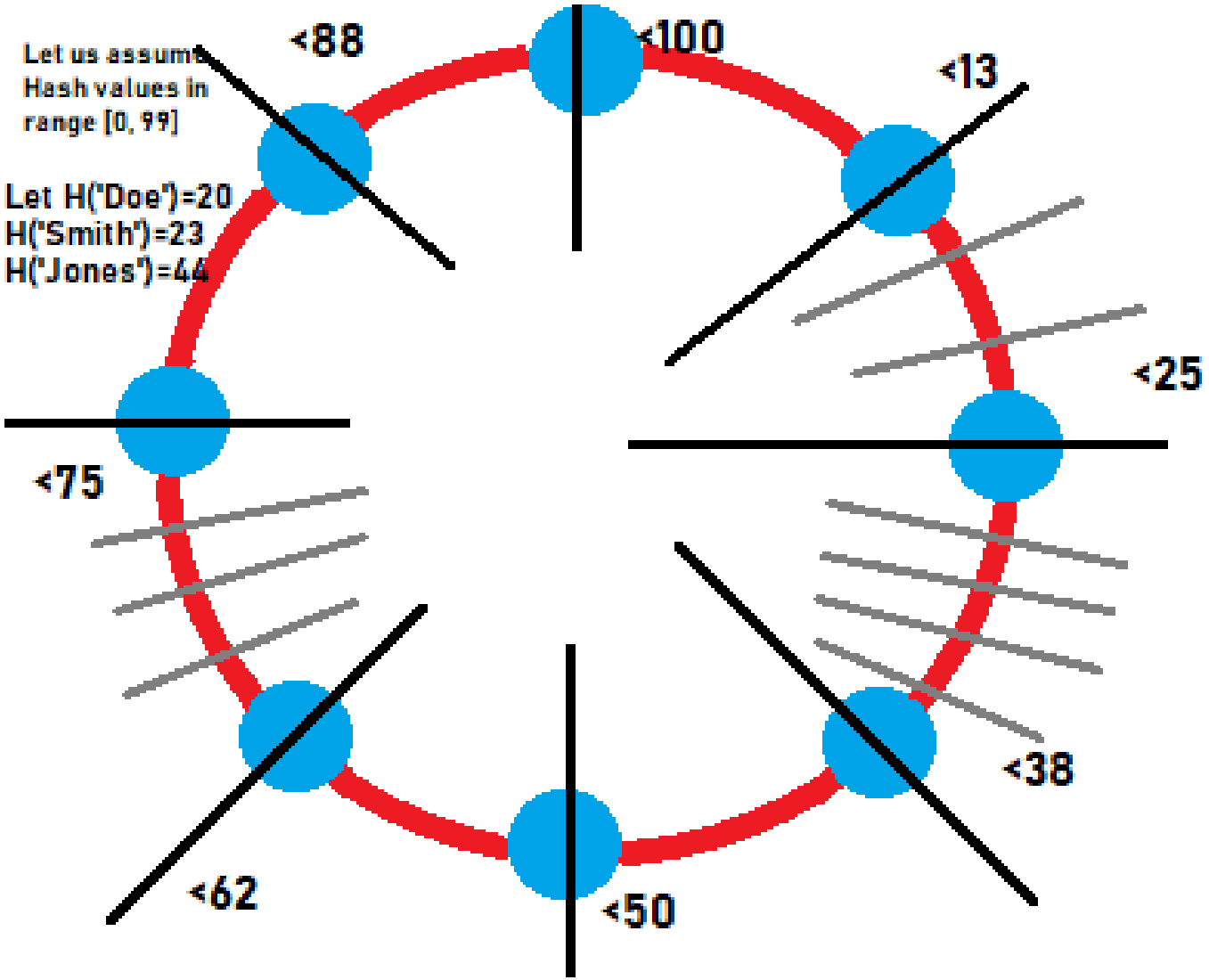
Ring Architecture



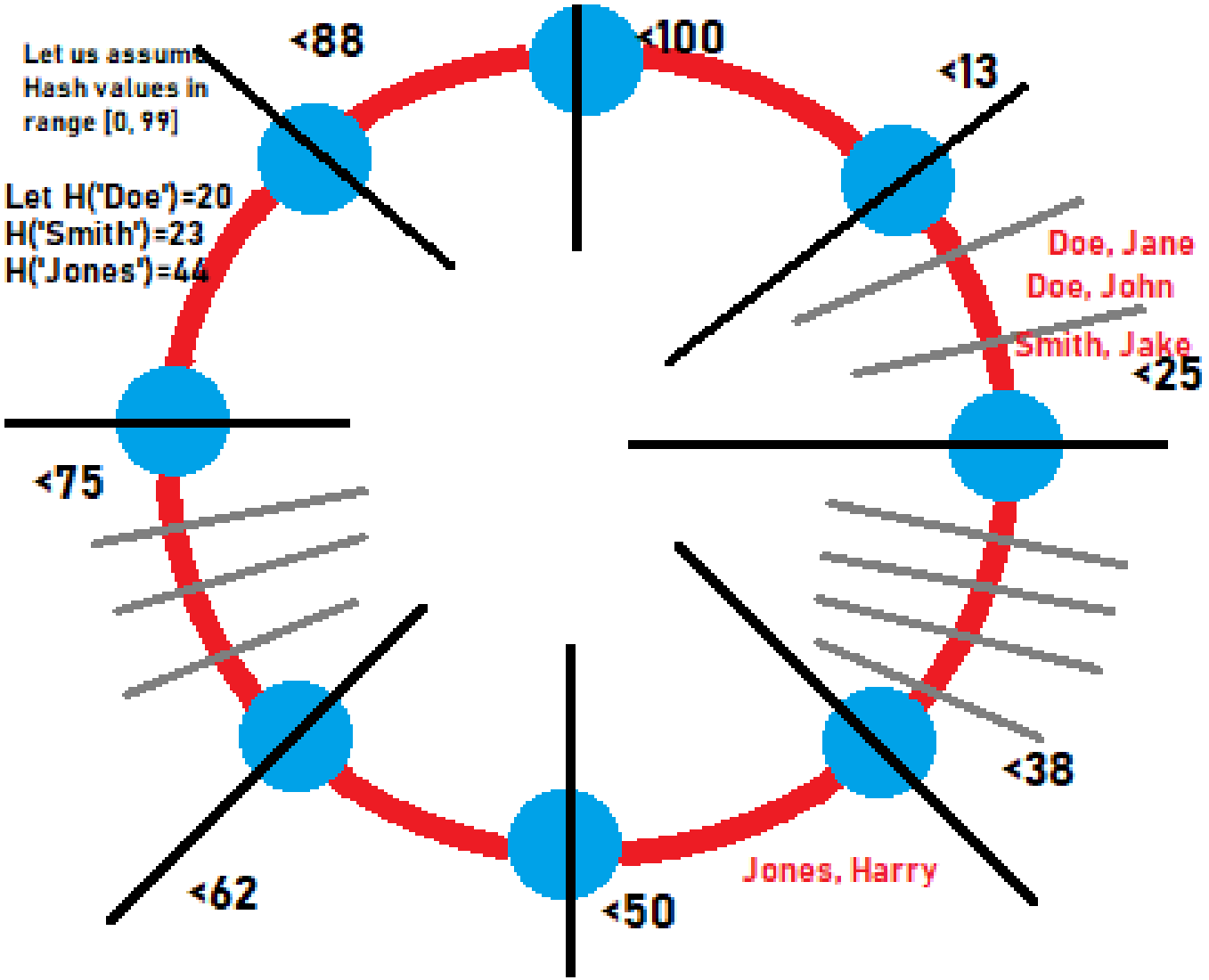
Ring Architecture



Ring Architecture



Ring Architecture



Ring Architecture

Organization of the nodes:

- physical nodes are organized into racks
- several racks may be connected (e.g., in the same localization)
- multiple racks, not necessarily directly connected, create a data center
- `nodetool status` displays information on nodes - their up or down status, rack and data center assignments
- the assignment for a specific rack and data center is stored in `cassandra-rackdc.properties` file

Replicas

- In order to prevent data loss on node failure and to speed up the response, data is stored in multiple nodes.
- Replication factor property specifies how many replicas we want to keep (by default 3).
- Basic replication strategies include: Simple strategy, local strategy and network topology strategy
- `SimpleStrategy` (recommended for single data center) – moved clockwise on the ring and decides the nodes at random
- `LocalStrategy` (is used only for internal keystores) – distributes among nodes in a single cluster
- `NetworkTopologyStrategy` – keeps data spread and allows to specify number of replicas kept in each data center

Replicas

- Since data is kept on multiple nodes, we have to assure consistency.
- Whenever data is requested a consistency parameter is read to determine the policy on the number of nodes that have to return the answer.
- Consistency levels:
  - `ANY` (returns successful write if at least one node is written)
  - `(local) <number>` e.g., `TWO` (returns when selected number of nodes are written to/read from – in case of different values, decides based on the newest and repairs older values)
  - `(local) QUORUM` (waits until majority responds, the newest value is selected and older are fixed)
- `CONSISTENCY` command checks current consistency level (it is set globally for a session), to change it – `CONSISTENCY <type>`, e.g., `CONSISTENCY QUORUM`

Snitches

- Snitches are responsible for determining which node is located in which rack and data center.
- A snitch assigns user for connection to a given node based on the query, their location and information about the delays of the nodes.
- Several snitch types are supported by Cassandra, including multiple vendor–specific ones.
- *Simple snitch* – when all nodes are on a single server
- *Property File Snitch* – each node has a file with IPs, data center and rack of each node
- *Gossiping Property File Snitch* – each node keeps only its own data center and rack information in `cassandra-rackdc.properties` file, information spread via gossiping protocol
- Snitch configuration is kept in `cassandra.yaml`
- A specific snitch type that works simultaneously with others is `Dynamic Snitch` that keeps track of the location of the replicas and calculates delays to those nodes

Basics

- Cassandra provides a simple query language that is based on SQL3.
- The data is stored vertically, inside of keyed tables that are kept in a keyspace.
- Each table has to have a primary key that consists of (possibly complex) partition key and optional clustering columns.
- Each column has a single (possibly complex) data type.
- In a row, inside a fixed non-key column, there can be stored an instance of a given data type, a `null` or nothing at all.
- Due to time-efficiency there are significant restrictions in the possibilities of a query (no JOINs, restricted WHERE and ORDER BY), hence tables are created with a specific query in mind, leading to data redundancy.

Basics

indexNo	lastName	firstName	major
200831	{lastName:Doe}	{firstName:Jane}	{major: BDA}
200832	{lastName:Doe}	{firstName:John}	{major: BDA}
200833	{lastName:Jonson}	{firstName:John}	{major: BDA}
200999	{lastName:Smith}	{firstName:Frank}	{major: BDA}
200840	{lastName:Smithy}	{firstName:Paul}	{major: CS}
220832	{lastName:West}	{firstName:Eve}	{major: OP}

Basics

indexNo	lastName	firstName	major
200831	{lastName:Doe}*2	{firstName:Jane}	{major: BDA}*4
200832		{firstName:John}*2	
200833	{lastName:Jonson}		
200999	{lastName:Smith}	{firstName:Frank}	
200840	{lastName:Smithy}	{firstName:Paul}	{major: CS}
220832	{lastName:West}	{firstName:Eve}	{major: OP}

Data types

Simple types:

- logical – Boolean
- numerical – (big)int, decimal, double, float, varint
- alphabetic – ascii, inet, text, varchar

Special types:

- blob
- counter
- timestamp
- uuid and timeuuid
- user defined types (`CREATE TYPE`, `ALTER TYPE`)

Collection types:

- list<type>
- set<type>
- map<type, type>

Data types

```
CREATE TYPE address (
city text,
postalCode text,
street text,
building int,
flat int,
phone set<int>
);

ALTER TYPE address ADD email text;
```

Data types

- uuid – a (pseudo)random, unique, identifier with a preset length, we can generate a new one using `uuid()`
- timeuuid – a (pseudo)random, unique, identifier with a preset length, that is partially dependent on the time of creation, i.e., IDs created after each other are similar and may be ordered, we can generate a new one using `now()`
- counter – a non-settable type, that may be modified only via updates, in particular cannot be initialized, we start with update column=column+1, so that counter column gets value 1, later on we increment or decrement the value

Keyspaces

- Keyspaces are containers that keep the data-filled tables.
- Since no JOINs are possible between tables, keyspaces are kept for managing the settings and parameters.
- `CREATE KEYSPACE` <name> <options> (e.g., `CREATE KEYSPACE` dbKS `WITH` replication={'class' : 'SimpleStrategy', 'replication_factor': '3'} `AND` 'durable_writes'='false');
- `DESCRIBE KEYSPACES` shows available keyspaces, `USE` dbKS; `DESCRIBE TABLES`; describes tables in a given keyspace
- To remove a keyspace we use `DROP` <ksName>.

Tables

- CQL uses tables similar tables to the ones known from relational databases.
- We can create a new table using `CREATE TABLE` <name> (field1 type1 `PRIMARY KEY`, field2 type2, ...);
- Each column has to be of a single data type.
- Check constraints and triggers are not supported.
- We change the tables using `ALTER TABLE`, we remove a table using `DROP TABLE`, a single command for deleting all rows in a table is also available – `TRUNCATE` <name>.
- We use `DESCRIBE` <tablename> in order to get information (columns, CREATE statement, compression parameters, Bloom filter FPR, TTL, index intervals, . . .)
- Key specification may include multicolumn keys – `CREATE TABLE` <name> (field1 type1, field2 type2, ..., `PRIMARY KEY`(<partition key>, <clustering column>));
- The partition key may be complex, e.g., (lastName, firstName) with the order determining the indexing.

Tables

Examples

```
CREATE TABLE StudendsByID (indexNo int PRIMARY KEY, lastName text, firstName text, major text);
```

```
CREATE TABLE StudendsByName (lastName text, firstName text, startingYear int, email text, indexNo int, major text, PRIMARY KEY((lastName, firstName), startingYear)) WITH bloom_filter_fp_chance = 0.03;
```

```
ALTER TABLE StudendsByName ADD tutor text;
```

```
DESCRIBE StudendsByName;
```

```
DROP TABLE StudendsByID;
```

Selects

- Selects follow the same syntax as in SQL however there are significant constraints.
- The select query, by default, may not condition the selection (no `WHERE` clause) or has to use the whole partition key in the condition.
- In order to allow conditions on a different column, we have to use `ALLOW FILTERING`, use a different table, a view or create a secondary index.
- We can use writetime(column) to get the time of the last update of each row in the column.
- The results may be `ORDER`ed BY one or many clustering columns.
- We can `GROUP BY` by one or more clustering column.
- We can use standard aggregate functions – `COUNT`, `SUM`, `MIN`, `MAX`, `AVG` or create our own.

Selects

Examples

`SELECT * FROM StudendsByName;`

`SELECT lastName, writetime(lastName) FROM StudendsByName;`

`SELECT * FROM StudendsByName WHERE lastName='Doe';` — will result in an error

`SELECT * FROM StudendsByName WHERE lastName='Doe' AND firstName='Jane';`

`SELECT * FROM StudendsByName ORDER BY startingYear;`

`SELECT MIN(startingYear) FROM StudendsByName;`

Creating aggregate functions

```
CREATE OR REPLACE FUNCTION avgState ( state tuple<int,bigint>, val int )
CALLED ON NULL INPUT
RETURNS tuple<int,bigint>
LANGUAGE java AS
$$ if (val !=null) {
      state.setInt(0, state.getInt(0)+1);
      state.setLong(1, state.getLong(1)+val.intValue());
      }
   return state; $$
;

CREATE OR REPLACE FUNCTION avgFinal ( state tuple<int,bigint> )
CALLED ON NULL INPUT
RETURNS double
LANGUAGE java AS
  $$ double r = 0;
     if (state.getInt(0) == 0) return null;
     r = state.getLong(1);
     r/= state.getInt(0);
     return Double.valueOf(r); $$
;

CREATE AGGREGATE mean(int)
SFUNC avgState
STYPE tuple<int,bigint>
FINALFUNC avgFinal
INITCOND (0,0);
```

Inserts

- Inserts follow a familiar syntax, yet we always should enumerate the columns we are inserting data into and we do not have to insert anything into non-key columns.
- If we insert data with the same primary key, the data is overwritten.
- `INSERT INTO` <table> (columns) `VALUES`
- `INSERT INTO` addressBook(name, email) `VALUES` ('John Doe', ['JD@gmail.com','john@doe.com']);
- `INSERT INTO` phoneBook(name, phone) `VALUES` ('John Doe', {555555555, 555123456});
- `INSERT INTO` location(name, address) `VALUES` ('John Doe', {'home': '81st Street, NY, NY', 'office' : 'Grunwaldzka, Wrocław' });

Updates

- Similarly to selects, updates require the whole primary key as a condition in `WHERE` clause.
- Cassandra supports temporary updates: `UPDATE ... USING TTL` <seconds> changes the value only for a given time, after the expiration it is removed (so we have no value in the column).
- We can not update simple-data-columns as in SQL and we can modify the contents of a collection – adding to set column = column + {new values}, removing by column = column - {old values}, column={} removes all values
- `UPDATE` phoneBook `SET` phone = phone + {555000001} `WHERE` name = 'John Doe';
- `UPDATE` location `SET` address = address-{'home': '81st Street, NY, NY'} `WHERE` name = 'John Doe';

Deletes

- Deletes follow the same rules that updates.
- We can delete the whole row, just like in SQL `DELETE FROM <table> WHERE <condition>;`
- We can also delete a specific column value in a single row (similarly to not inserting the value it is displayed as null, yet nothing is stored) `DELETE <column> FROM <table> WHERE <condition>;` Naturally, all the restrictions on the deleted columns (e.g., cannot be a part of the key) and conditions apply.
- TTL removal after the time passes is not a delete and is bascially free operation in terms of cost.

Batches

- As a distributed database Cassandra does not support ACID transactions.
- We have a counterpart – `BATCH` operations.
- `BATCH`es allow us to use multiple modification queries (inserts, updates, deletes), possibly on a different tables, as an (almost) atomic operation.
- Syntax:

```
BEGIN BATCH
<inserts>
<updates>
<deletes>
APPLY BATCH;
```